Patterns of correlation within the human methylome

Brook O'Reilly

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Biological Sciences

University of Canterbury

2021

Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance.

I would like to thank my primary supervisor, Dr. Amy Osborne, for accepting me as a Master's student and mentoring me in a field in which I initially had very little experience. Without her judgement, this work would have been drastically different. I would also like to thank my secondary supervisor, Dr. Miles Benton, for his valuable expertise and encouragement.

I would like to thank all of the researchers who published their research freely online, and the institutions and other members of the community which opted to make these publications open-access, for their contributions to the accessibility of scientific knowledge in the face of an increasingly profit-driven system.

Finally, I'd like to thank everyone who has contributed to the open-source software used throughout this thesis. Technological process would be significantly slower if not for our ability to 'stand on the shoulders of giants' and this applies as much to those who develop the tools as it does to those who used them.

This work was made possible by the use of the RCC facilities at the University of Canterbury.

Abstract

Traditionally, epigenome-wide association studies have focused on the relationship between epigenetic marks and specific traits by relating the methylation intensity of CpG sites with a phenotype. Rather than interpreting these values and relate them to the expression of a gene, our research investigates the correlation in DNA methylation intensity between CpG sites on the same chromosome within the human genome. We postulate that a strong correlation in methylation intensity (as we have defined) can be indicative of an underlying epigenetic trend; an analysis of correlation only requires DNA methylation data, so we can attempt to identify underlying these epigenetic associations in the absence of gene expression information (such as that obtained through RNA sequencing or recording or phenotypes).

We make use of existing DNA methylation array data to answer our research questions:

- What constitutes a meaningful strong correlation in methylation intensity?
- Does array normalisation have an effect on the strong correlations we can detect?
- Is correlation strength influenced by the distance between a pair of CpG sites?
- Does co-location of two CpG sites within a functional group such as a gene or pathway tend to produce a stronger correlation in methylation intensity?
- Are we able to identify new biological pathways (or support existing ones) by looking at correlations in methylation intensity?

Our research identified that the strongest 10% positive and negative Spearman correlation coefficients were a suitable subset for our analyses. Array normalisation was shown to have a profound effect on methylation intensity, and there was some impact on correlation as well. Correlation strength did not follow a linear relationship with distance between the correlating pair, though we did see that CpG sites within the same gene tended to correlate more strongly. Our research suggested that it may be possible to use correlation analysis to identify pathways in some circumstances, though future research is needed to develop a robust approach for this. An important incidental finding was that CpG islands tend to contain much stronger correlations in methylation intensity than the average for a chromosome - a fact which may be exploitable for future delineation of CpG islands.

Contents

0.1 Abbreviations

5mC	5-methylcytosine
BCM	Beta Correlation Matrix
BER	Base Excision Repair
CHDS	Christchurch Health and Development Study
CpG	Cytosine (phosphate) Guanine
CSV	Comma-separated values, a file format
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
EWAS	Epigenome-wide association study
FunNorm	Functional normalisation
HPC	High-performance computing
MBD	Methyl-CpG-binding domain
NOOB	Normal-exponential using out-of-band probes
Quant	Stratified quantile normalisation
SNP	Single-nucleotide polymorphism
SWAN	Subset-quantile within array normalisation
TAD	Topologically-associating domain
TET	Ten-eleven translocation

Part I

Introduction

Chapter 1

Overview

1.1 Background

Epigenetics is the study of the stable and heritable alterations of gene expression that occur during cellular development and proliferation. Unlike conventional genetics, which is more concerned with the implications of an organism's genome sequence, epigenetics looks at non-sequence changes that are associated with variability in gene expression.

There are a number of known epigenetic regulation mechanisms, including histone modification (Bannister and Kouzarides, 2011), RNA interference (Hannon, 2002) and DNA methylation (Moore et al. 2013). This study will focus on cytosine-based DNA methylation - the process by which methyl groups are attached to cytosine residues in DNA, forming 5-methylcytosine. The presence of these methyl groups can regulate gene expression by modulating the interaction between DNA and proteins involved in transcription (Jaenisch and Bird, 2003). More specifically, it has been shown that the methylation level of the first intron of a gene inversely correlates with gene expression in mammals (Anastasiadi et al. 2018).

A family of enzymes known as DNA methyltransferases (DNMTs) catalyse the DNA methylation reaction (Moore et al. 2013), which is reversible via enzymes such as those from the ten-eleven translocation (TET) family (Kohli and Zhang, 2013). The ability for DNA to be reversibly methylated allows an organism to better-respond to the environment by changing gene expression, though it can also occur stochastically as an organism ages (Jaenisch and Bird, 2003). Epigenetic alterations are known to be heritable, both in the context of cellular proliferation and at the whole-organism reproductive level (Trerotola et al. 2015).

Epigenetic alterations are known to occur in response to external or environmental factors and have been shown to play a role in a number of major developmental events (Jaenisch and Bird, 2003; Kiefer, 2007). The environmental response provides a mechanism wherein gene expression levels can be propagated from one generation of cells to the next, boosting an organism's ability to survive stress conditions in the mid- to long- term. The reversibility of epigenetic alterations provides the advantage of being able to adapt to a changing environment without relying on mutation or other genomic sequence alterations (Jaenisch and Bird, 2003). In terms of organism development, DNA methylation is known to silence gene expression in the long-term as appropriate for each cell type – epigenetic expression has been described as the 'software' of the

genome that directs embryogenesis and development, as opposed to the 'hardware' comprised of the genome itself (Kanherkar et al. 2014). DNA methylation is passed between whole-organism generations via methods such as genomic imprinting (Reik et al. 2001) and the epigenetic consequences of stressors such as major famine have been shown to be trans-generational (Veenendaal et al. 2013). Section 1.2 reviews a number of important considerations related to DNA methylation which provide justification for the hypotheses of this study, which are described in section 1.6.

1.2 Overview of DNA methylation

The conversion of cytosine to 5-methylcytosine (5mC), as well as the reverse reaction, comprise the most well-studied mechanism of methylation-based epigenetic regulation. Though the existence of 5mC in the mammalian genome was known as early as the 1940s (Hotchkiss, 1948), its role in mammalian gene regulation was not shown experimentally until the 1980s. Compere and Palmiter (1981) showed that synthesis of a specific mRNA in a mammalian cell line correlated with hypomethylation of its associated gene, and that this change in expression was passed between cells as an epigenetic change, i.e. the expression change was retained in daughter cells despite no apparent alteration of genetic sequence. More generally, DNA methylation alters gene expression by directly inhibiting binding of transcription factors (Brenet et al. 2011) or by recruiting proteins with other epigenetic effects, such as methyl-CpG-binding domain (MBD) proteins which are involved in histone modification (Roloff et al. 2003). As gene expression is inextricably linked to phenotype, DNA methylation is known to have a significant role in cell development (Smith and Meissner, 2013) and disease etiology (Robertson and Wolffe, 2000).

In mammals, DNA methylation occurs primarily at CpG sites, which are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in its 5' to 3' direction. Assuming that the distribution of bases in DNA is uniform (i.e. they all appear at roughly the same rate) then the probability that any two consecutive bases are a C and G would be about 1 in 16, or 6.25%, equating to about 194 million CpG sites throughout the human genome. However, there are only about 29.3 million CpG sites observed in recent assemblies of the human genome (Luo et al. 2014), or just under 1% - drastically lower than what one would assume from a probabilistic standpoint. In unmethylated DNA, spontaneous deamination of cytosine results in the creation of uracil, which is typically repaired via the uracil-DNA glycosylase pathway. The methylated form, 5-methylcytosine, deaminates into thymine and remains as a permanent mutation until repaired by the BER pathway (Duncan and Miller, 1980; Krokan and Bjørås, 2013). Consequently, CpG sites are fewer in number than one would expect. Despite the relatively low abundance of CpG dinucleotides in the human genome, there are regions that show a dramatic increase in CpG frequency, referred to as CpG islands (Larsen et al. 1992). These regions are often associated with gene promoters (Saxonov et al. 2006; Illingworth et al. 2010), and their methylation is considered to be play a major role in gene silencing (Illingworth and Bird, 2009). About half of all identified CpG islands contain transcription start sites, with the other half being referred to as 'orphan' CpG islands owing to the uncertainty of their purpose (Deaton and Bird, 2011).

The procedure of 'writing' DNA methylation is carried out by a family of enzymes known as DNA methyltransferases, or DNMTs, of which there are many different types (Lyko, 2018). The specifics of each DNMT are beyond the scope of this thesis, but DNMTs can generally be assigned to one of two categories (Moore et al. 2013):

- *de novo* methyltransferases, which in concert with cofactors, catalyse the methylation of cytosines. DNMT3A is one such methyltransferase and is associated with genomic imprinting (Kaneda et al. 2004)
- maintenance methyltransferases, which maintain existing patterns of DNA methylation. DNMT1, for example, is associated with the epigenetic maintenance of haematopoietic stem cells (Trowbridge et al. 2009)

'Erasing' DNA methylation can occur passively or actively. Maintenance errors or inhibition during cell division can result in lowered levels of DNA methylation (Moore et al. 2013), and as described previously in this section, 5mC can be spontaneously deaminated and repaired into thymine which causes a G/T mismatch - repair via the BER pathway would not re-implement the methyl group. Active DNA methylation, such as that which occurs during epigenetic reprogramming (see section 1.2.4) occurs due to enzymes such as the ten-eleven translocation (TET) methylcytosine dioxygenases and their associated pathways (Wu and Zhang, 2017).

Recent studies have shown that bases other than cytosine can also be methylated in mammals in both DNA and RNA. Methylation of adenine to form N^6-methyladenine (6mA) is a common DNA modification in prokaryotes, but was largely ignored in eukaryotes up until recently (reviewed by Heyn and Esteller, 2015). A number of studies have suggested that 6mA plays a significant role in human epigenetic regulation - some examples of this include work by Wu et al. (2016) suggesting adenine methylation plays a role in embryonic development; and a paper by Xiao et al. (2018) suggesting that levels of 6mA influence human tumorigenesis. Conversely, an article by Douvlataniotis et al. in March 2020 asserts that current approaches to assessing 6mA in mammals are flawed and publications they reviewed (including those by Wu et al. and Xiao et al.) provide insufficient evidence to support the action of 6mA as significant epigenetic mechanism in mammals. This thesis focuses on cytosine methylation data, but many of the concepts and hypotheses may still apply and could be a potential avenue for future research.

1.2.1 Environmental influences

The conditions that an individual exposes themselves to on a regular basis are often referred to as 'lifestyle factors' and can have a significant impact on DNA methylation. Diet and nutrition are also known to greatly influence DNA methylation in specific tissue types (Ulrey et al. 2005; Kadayifci et al. 2018) - a recent study proposed that changes in DNA methylation may play a role in the effectiveness of the ketogenic diet in the treatment of epilepsy (Chen et al. 2019). Exercise is associated with altered regulation of a number of genes typically associated with good health (Sailani et al. 2019; Ferrari et al. 2019). It has also been shown that a particularly sedentary lifestyle has effects on DNA methylation (van Roekel et al. 2020). Sleep deprivation, a condition affecting millions of people across the world, has significant impact on the epigenome as a whole (Gaine et al. 2018) and DNA methylation specifically is implicated in the glucose-managed maintenance of the circadian rhythm (Peng et al. 2019).

Environment in the more-literal sense also has an impact on the epigenome. Temperature and humidity have been shown to have an effect on DNA methylation (Bind et al. 2014), as has sun exposure (Grönniger et al. 2010). Environmental exposure to pollutants has been shown to alter DNA methylation, especially at younger ages; cellular toxicity due to the presence of arsenic has been shown to impact the methylome in

multiple different age groups (Bandyopadhyay et al, 2016; Hossain et al. 2017; Lambrou et al. 2012) as well as in utero (Kile et al. 2012). Tobacco smoke has similarly been shown to change DNA methylation across all age ranges, particularly in utero (Joubert et al. 2012; Peluso et al. 2014) and the effects remain over the long-term (Shenker et al. 2013). Martin and Fry (2018) review a number of studies related to the effects of environmental pollutants on the human epigenome.

1.2.2 Age influences

The association between an individual's chronological age and their DNA methylation patterns has been studied in depth for some time (Boks et. al 2009; Teschendorff et. al 2010; Rakyan et. al 2010). A surge of academic interest in the epigenetics of aging during the early 21st century quickly lead to researchers considering the prospect of using DNA methylation as a predictor of age and age-related diseases. A seminal paper by Horvath et. al in 2011 was followed shortly thereafter by a number of estimators for chronological or biological age which use methylation status of particular CpG sites within a given tissue type (Horvath, 2013; Hannum et. al, 2013; Weidner et. al, 2014; Giuliani et. al, 2016). The difference between an individual's chronological age and their age as predicted by one of these 'clocks' is referred to as epigenetic age acceleration (Horvath and Raj, 2018). This acceleration can result in someone being epigenetically younger or older than their chronological ages, which comes with its associated age-associated morbidities (Horvath et al. 2015, Marioni et al. 2015). Individuals with Werner syndrome, a disease characterised by the appearance of premature aging, may be subjected to faster epigenetic aging as per a study which compared the chronological age of individuals with their DNA methylation-based clock age (Maierhofer et al. 2017).

Two key observations can be made with regards to the effects of aging on the methylome: genome-wide hypomethylation, and promoter-specific hypermethylation (Bollati et al. 2009; Johnson et al. 2012). The rate of passive change of DNA methylation, referred to as part of a broader 'epigenetic drift', is stochastic in nature resulting in compounding of DNA methylation maintenance errors over time (Issa, 2014). Consequently, these effects can vary greatly between even genetically-identical individuals. Multiple twin studies have found that while monozygotic twins are epigenetically very similar when they are young, older twins can differ significantly in their DNA methylation (Fraga et al. 2005; Talens et al. 2012). It has been suggested that epigenetic mosaicism occurring in-part as a result of DNA methylation drift is one of the key factors that lead to phenotypes associated with aging (Issa, 2014).

1.2.3 Genetic influences

Epigenetics, literally 'on top' of genetics is ultimately subject to an individual's genomic sequence. Genetic variation will result in epigenetic variation to some degree. For example, single nucleotide polymorphisms at a CpG site may remove the site completely, preventing cytosine methylation. There has been a number of studies researching the effects of SNPs on DNA methylation of nearby CpG sites (Soto-Ramírez et al. 2013; Veenstra et al. 2018). The effects of genetic variation on the methylation status of CpG sites of interest throughout the genome have also been studied (Gaunt et al. 2016) - this study presented a catalogue of genetic influences on DNA methylation in human blood at a number of different life stages, suggesting that the genetic component of methylation may have a causal role in complex traits.

Possibly the most striking genetic influence on the mammalian epigenome is caused by the presence of

multiple X chromosomes, as is typically the case in females. The phenomenon of X-inactivation and its role in mammalian development was first described by Mary Lyon (1972). Essentially, one of the two X chromosomes is 'inactivated' at random to prevent excessive expression of genes located on this chromosome (Brockdorff and Turner, 2015). This is associated with a number of epigenetic effects - most relevant to this thesis is the impact on DNA methylation, which can be a tendency to increase or decrease in a site-dependent manner (Sharp et al. 2011).

1.2.4 Inherited influences

One of the key attributes of epigenetic modifications is that they are heritable to some extent. Within a few hours of fertilisation, DNA methylation patterns in the newly-formed zygote are largely erased or 'reset' as part of a process referred to as 'epigenetic reprogramming' (Fraser and Lin, 2016) though some marks present in the parental methylomes are retained (Wang et al. 2014). The expression of certain genes changes depending on whether that gene was inherited from the mother or father, owing to a process referred to as 'imprinting' (Reik and Walter, 2001; Ferguson-Smith, 2011). DNA methylation plays a significant role in mammalian genomic imprinting, and the epigenetic marks on many imprinted genes are retained after epigenetic reprogramming (Bartolomei 2009; Bartolomei and Ferguson-Smith, 2011).

Additionally, stresses on the either parent prior to fertilisation, or during pregnancy in the case of the mother, can manifest as an epigenetic influence. A well-studied example of epigenetic inheritance from the father is the 1944-45 Dutch famine which has been the subject of a number of trans-generational studies. It has been shown that not only the direct descendants who were prenatally exposed to this famine via their parents had a famine-influenced DNA methylation pattern (Heijmans et al. 2008) but children of prenatally under-nourished fathers (but not mothers) were comparatively heavier and more obese than children whose fathers were not under-nourished (Veenendaal et al. 2013). As an example of maternal epigenetic inheritance, a study found that prenatal maternal hardship due to the 1998 North American ice storm correlated with a deviation of methylation levels of a large number of genes associated with immune function in their children (Cao-Lei et al. 2014).

1.2.5 Tissue type

The study of epigenetics has answered many of the questions relating genotype to phenotype and provides a justification for how a single genome can produce many different cell types. Studies have shown that patterns of DNA methylation (and changes thereof) are dependent on the local tissue type (Thompson et al. 2010; Horvath 2013), and a number of initiatives exist to capture this information, such as the NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al. 2010).

The study of correlations in samples derived from tissues other than whole blood is beyond the scope of this thesis, though this presents a promising area of future research.

1.2.6 Dysregulation of DNA methylation

As with other epigenetic phenomena, DNA methylation is ultimately a mechanism which controls cellular development and metabolism via regulation of genes relevant to these processes. As described in section 1.2, DNA methylation alters gene expression by directly inhibiting binding of transcription factors or by

recruiting proteins with other epigenetic effects. Dysregulation of this process results in sub-optimal gene expression and can often lead to disease phenotypes.

Several genetic disorders that impact DNA methylation have been characterised. ICF syndrome, named for its presentation of immunodeficiency, centromeric instability and facial anomalies, is an extremely rare genetic disorder (Brown et al. 1995). An epigenetic characteristic of ICF syndrome is a genome-wide loss of DNA methylation owing to a mutation in DNMT3b, one of the DNA methyltransferases thought to be involved in de novo methylation (Heyn et al . 2012). Earlier research suggested that immunodeficiency occurring as a result of ICF syndrome may be a result of non-specific hypomethylation in regions critical to B cell development (Hansen et al. 1999). Rett syndrome is a genetic disorder characterised mainly by neurological disorders (in particular, seizures and apraxia) and reduced physical growth, primarily of the head (Smeets et al. 2012). The syndrome is caused by mutations in the MECP2 gene (Amir et al. 1999). As the inheritance pattern is X-linked, this mutation is lethal to the vast majority of genetic males who inherit it; consequently, Rett syndrome is predominantly associated with females (Chahil et al. 2018). MECP2, or methyl CpG binding protein 2 is an abundant nuclear protein which binds to methylated CpG sites in the genome and recruits transcriptional repressors such as mSin3A and histone deacetylases (Nan et al. 1998). A more recent study has shown that it can also activate transcription, rather an suppress it, and postulates that aberrant MECP2-facilitated regulation of genes in the hypothalamus may be the main driver behind Rett syndrome (Chahrour et al. 2008). Fragile X syndrome is another X-linked dominant genetic disorder characterised by intellectual disability (Bagni et al. 2012), though it is significantly less lethal; one study suggested that the average age of death was only about 12 years lower than the general population but admits that this may have been biased (Partington et al. 1992). Fragile X syndrome occurs as a result of a mutation of the FMR1 gene which causes an increase in CGG trinucleotide repeats in the 5' untranslated region - this leads to hypermethylation (and deacetylation) of the gene which suppresses transcription (Crawford et al. 2001).

In terms of cancers, DNA methylation has been proposed as an early diagnostic biomarker for several types (Dong et al. 2014) and the etiology of many different cancers involves aberrant regulation of tumour suppressor genes (Dong et al. 2014; Paska et al. 2015; Tse et al. 2017; Saghafinia et al. 2018; Sun et al. 2018; Wang et al. 2020). Dysregulation of genes required for normal development, such as the HOX genes, is also associated with oncogenesis despite these genes having no apparent onco-suppressive ability; rather, these proto-oncogenes only result in disease phenotypes if their epigenetically-driven expression is incorrect (Bhatlekar et al. 2014). Some studies suggest that aberrant DNA methylation is the cause of cancer and could be a potential target for medical intervention (Esteller, 2005; Mossman and Scott, 2006) - researchers have proposed the 'correction of a diseased epigenome' as a potential treatment or cure for cancer (Takeshima et al. 2019). Indeed, some clinical trials of epigenetic treatments for cancer have been conducted or are currently underway (Cheng et al. 2019).

An individual's age is by far the strongest indicator of mortality risk - the older one gets, the more likely they are to succumb to any number of diseases (Rae et al. 2010). In terms of DNA methylation, the trend with age is that hypomethylation tends to occur in heterochromatic regions of the genome, particularly in repetitive elements (Bollati et al. 2009; Heyn et al. 2012), though age-associated hypermethylation has also been shown to occur (Madrigano et al. 2012). As discussed in section 1.2.2, researchers such as Horvath

and Hannum have developed epigenetic clocks that can predict one's age based on specific CpG sites within their methylome to a high degree of accuracy. Some researchers have suggested that aging is the result of the accumulation of epigenetic modifications over time (Kane and Sinclair, 2019). As a consequence, researchers have invested some effort into reversing these epigenetic modifications in an attempt to find a therapeutic treatment for the consequences of aging. An experiment in an animal model showed that its possible to restore a 'youthful' epigenetic state which improves tissue function in aged mice (Lu et al. 2020). The idea of a youthful versus an elderly epigenetic state ties into the earlier discussion regarding cancer. Cancer rates increase exponentially with age (Cancer Research UK, 2020) and is also thought to be driven in part by DNA methylation. Multiple studies have proposed that cancer and aging are inextricably linked to the accumulation of aberrant patterns of DNA methylation, or 'epigenetic legions', over the course of one's lifespan (Esteller, 2000; Liu et al. 2003; Daniel et al. 2015; Unnikrishnan et al. 2018). Research in the field is progressing at a rapid pace, and the concept of an epigenetic treatment for both aging and age-associated diseases such as cancer may prove itself to be a reality in years to come.

1.3 Correlations and their applicability to DNA methylation

Correlation between a pair of variables suggests that there may be an underlying association between the variables. In general terms, a correlation is only a statistical association within data and does not mean that a causal relationship exists within that data.

In the context of DNA methylation, we calculate correlation between the degree of methylation present at individual CpG sites. We use beta values for this calculation as they are a linear representation of methylation intensity, bound between 0.0 (predominantly unmethylated) and 1.0 (predominantly methylated). A correlation in this case suggests that as one particular CpG site increases in methylation intensity, another specific CpG site will also tend to increase in methylation intensity (or decrease in the case of a negative correlation).

While our ability to accurately measure DNA methylation has increased dramatically over the past few decades, it is still an imperfect science and we would expect some noise and technical bias regardless of sequencing platform. We can attempt to reduce this using statistical approaches such as using sufficiently-large datasets and making informed choices when it comes to array normalisation. Assuming these were sufficiently accounted for, and taking only sufficiently-strong and statistically-significant correlations for our analysis, we would have a set of CpG pairs for which one of the following would be true:

1. Both CpGs are influenced by a common underlying epigenetic mechanism
2. Methylation of one CpG directly influences methylation of the other

These are discussed in more detail in section 1.3.3.

Other EWAS tend to focus on methylation intensity values directly, rather than focusing on associations. The use of different sequencing platforms and processing methods adds a layer of difficulty when comparing studies, but we would expect biologically-grounded, strong correlations to remain in place regardless. It follows that we may be able to use a correlation-based approach to identify biological pathways that are regulated (at least in part) by DNA methylation.

1.3.1 Limitations of correlations

1.3.1.1 Representation and minimum sample size

Statistically-speaking, a cohort will tend to have characteristics that are an approximation of the overall population. These approximations get closer to the overall population characteristics as the sample size grows larger, but unless the entire population is represented within that sample, there is generally going to be some divergence in any particular sample. In the context of what we are doing here, the statistical properties of any given cohort, such as the distribution of methylation intensities at any given CpG site, are only going to approximate the human population at large. An excellent p-value (or other metric) for a correlation coefficient means significantly less if the underlying distribution of the data is not a good representation of the population. We are a long way off being able to accurately assess the methylome of every human on the planet and the relative youth of the field of epigenetics (and in particular, population epigenetics) means that we do not have enough data to say for sure if any particular cohort is an accurate representation of our species as a whole.

To my understanding, no serious work on genome-wide methylation correlation has been published before. This is unfortunate, given the scientific potential of such a study, and compounds the issue mentioned above as we do not have any basis on which to identify a 'good' cohort for any correlation study. We do not know the minimum sample size required for decent and stable correlation computation. A past study has suggested that stable estimates of correlation occur as the sample size approaches 250 (Schönbrodt and Perugini, 2013) but it is not known how well this research applies to correlations in DNA methylation intensity, or even if it applies to all correlation methods - the authors did not explicitly mention which method they used, and their simulated dataset used randomly-generated values from a bi-variate Gaussian distribution, which may not be an accurate representation of the underlying distribution of the data we are using in this thesis.

1.3.1.2 Spurious correlation of ratios

A mathematical quirk often overlooked in the biological sciences is referred to as 'spurious correlation of ratios'. This was first described by Karl Pearson in 1897. The basic idea is that given three random uncorrelated variables x, y and z, a correlation between $\frac{x}{z}$ and $\frac{y}{z}$ will be found. This is because the pairs with a relatively large value of z will tend to be larger than those with a smaller one, and a significant correlation may occur between the two ratio variables as a result.

This study primarily uses beta values to represent methylation intensity. At a technical level, the beta value is a ratio representing the proportion of methylated reads at a given CpG site. This is an estimation of the probability that a CpG site is methylated, as opposed to a genuine ratio, so the principle of spurious correlation of ratios does not apply here. However, care must be taken to ensure other correlations between ratios are not misinterpreted in other studies.

1.3.1.3 Correlations due to randomness

Similar to the spurious correlation of ratios, we can occasionally see correlations in random uncorrelated data. Randomly-distributed pairs would have a correlation coefficient that tends towards zero for a sufficiently large sample size. For smaller datasets, we would expect these stochastic factors to manifest as a bell

curve in the overall distribution of correlations.

Deconvoluting 'legitimate' correlations from random ones is beyond the scope of this thesis. We can, however, make efforts to minimise the number of random-chance correlations by using only correlations above a certain threshold. Legitimate strong correlations will tend to be distributed closer to the extremes of 1.0 and -1.0 - the ratio of legitimate to random correlations (and therefore the likelihood that a given correlation is legitimate) increases for values closer to these extremes, so by setting a high threshold for a correlation to be considered 'strong', we also decrease the number of random-chance correlations in our set of identified strong correlations. There will always be some present with this method, but we will still be able to reduce the number (albeit at the cost of potentially losing some information regarding legitimately strong correlations).

1.3.2 Options for evaluating correlation

We can consider three common measures of correlation for the purposes of our studies:

- Pearson's Correlation
- Spearman's Rank Correlation
- Kendall's Rank Correlation

All three of these are compared in a brief study documented in chapter 5, and this section contains an overview.

An important assumption we make for the purposes of our research is that any statistical relationship between two CpG sites is monotonic but not necessarily linear, i.e. as one increases, the other will tend to increase or decrease, but not necessarily in proportion.

1.3.2.1 Pearson's Correlation Coefficient

The simplest canonical method of calculating correlation is by taking the Pearson correlation coefficient. This is the standard measure of correlation between two variables, and generally what most people think of when they think of correlation. Given a cohort of size n and paired data x_i and y_i with respective sample means of $\bar{x}$ and $\bar{y}$, the sample Pearson correlation coefficient r can be calculated with the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{1.1}$$

The Pearson correlation coefficient assumes a linear monotonic relationship between two variables. While we assume that the relationship between two associated CpG sites is monotonic, we don't assume that this relationship is necessarily linear; therefore, Pearson's method may be unsuitable for our research.

1.3.2.2 Spearman's Rank Correlation Coefficient

Spearman's rank correlation is a measure of how well a relationship between two variables can be described using an arbitrary monotonic function - in other words, whether or not a relationship exists between two variables that can be described as either:

1. As the value of one variable increases, the other variable will also tend to increase (positive correlation)

2. As the value of one variable increases, the other variable will tend to decrease (negative correlation)

Given the pairwise distances of the ranks of the variables d and a number of samples n, the Spearman correlation coefficient ρ can be calculated with the following equation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{1.2}$$

The Spearman rank correlation coefficient is usually larger in magnitude than the alternative Kendall coefficient. As a consequence, this may increase variance within our correlation data, as a tendency for larger magnitudes will cause the distribution of correlation coefficients to 'spread out'. This would make it easier to identify a threshold for selecting strong correlations as the frequency of correlations for a given range would be lower, as the statistical distribution would generally be less dense.

1.3.2.3 Kendall's Rank Correlation Coefficient

Kendall's rank correlation is similar to Spearman's in the sense that it is also a measure of how well a relationship between two variables can be described using an arbitrary monotonic function. The key difference is that it is based on concordant and discordant pairs; given a set of observations $(x_1, y_1), ..., (x_n, y_n)$, any pair of observations (x_i, y_i) and (x_j, y_j) are concordant if both $x_i < x_j$ and $y_i < y_j$, otherwise they are discordant. Given n_c concordant pairs and n_d discordant pairs, for a total of n pairs overall, the Kendall correlation coefficient τ can be calculated using the following equation:

$$\tau = \frac{2(n_c - n_d)}{n(n-1)} \tag{1.3}$$

Kendall's approach is the preferred method for calculating rank correlation for a small sample size as it provides better estimates than the Spearman method. Additionally, Kendall's rank correlation coefficient is easier to interpret, as it is directly linked to the probabilities of observing the concordant and discordant pairs.

1.3.2.4 Selection of method for preliminary studies

Given the assumption of a monotonic and possibly non-linear relationship, a rank correlation coefficient would likely be a better choice. A 2010 study by Croux and Dehon suggested that both Spearman and Kendall correlation are statistically 'robust' and 'efficient' measures, with the Kendall method being preferable due to its superior performance under the conditions of their study. An informal test (similar to the one discussed in section 5.2) found that computation of the Spearman correlation matrix was orders of magnitude faster than that of the Kendall method, using the Python packages available for our research. Given the need for evaluation of a number of large correlation matrices, and limitations on time available for computing, the Spearman measure will be used for preliminary studies.

1.3.3 Biological interpretation of correlations

As per section 1.3, we have identified two possible explanations for a significant correlation in DNA methylation intensity between two CpG sites.

The first explanation is that both CpGs are influenced by a common underlying epigenetic mechanism.

This is known in statistics as a spurious relationship - one wherein two variables are not causally related, but may be associated due to the presence of confounding factors. A positive correlation in this case could suggest that two distinct CpG sites both experience DNA methylation in response to the same environmental stimulus. A possible example of this could arise in genes associated with diet, where multiple correlating CpG sites are methylated or demethylated based on consumption of specific foods, as has been seen in fruit and fruit juice (Nicodemus-Johnson and Sinnott, 2017).

The second explanation is that methylation of one CpG site directly influences methylation of the other. In this case, a positive correlation could suggest that the presence of one gene product activates a pathway which results in the demethylation of CpG sites on another gene. A similar mechanism could exist in the reverse direction, wherein the absence of one gene product activates a pathway which results in the methylation of CpG sites on another gene. In the instance of a negative correlation, a negative feedback loop could exist, where an increase in one gene product activates a pathway which methylates another CpG site (perhaps to reduce its associated gene product), and vice versa.

The lack of studies so far into patterns of correlation in DNA methylation means that these two explanations are simply conjecture at this point. Nonetheless, they provide the basis for hypotheses and a justification for continuing with this research.

1.3.4 How correlations are used in this study

For array-derived data, the beta values for each CpG site are calculated individually for each individual (or distinct sample) in the cohort. The number of beta values for each CpG site is equal to the number of individuals (or distinct samples; both correspond to n) in the cohort. Any two CpG sites will have n data points consisting of two beta values, allowing a correlation coefficient to be calculated between the CpG sites for the cohort in general.

Due to the absence of past correlation studies in this area, it is difficult to define a 'strong' correlation in the context of DNA methylation intensity. One could select a threshold based on personal preference and go with that - for example, 0.7 - and any correlation with a magnitude greater than or equal to this is considered a strong correlation. Perhaps the better thing to do would be to take a certain percentage of the strongest positive and negative correlations and look for trends in this subset, rather than arbitrarily setting a threshold. Section 2.2.5 discusses our options for identifying strong correlations.

1.3.4.1 Correlation networks

We can define a correlation network as a set of pairs of CpG sites that are associated by means of at least one shared strong correlation in methylation intensity. In other words, a correlation network is an undirected graph with CpG sites as the vertices, with edges comprised by strong correlations between beta values of these CpG sites. The smallest possible correlation network is one consisting of two CpG sites that only strongly correlate with each other, i.e. they do not strongly correlate with any other site. Larger correlation networks, consisting of CpG sites that correlate with multiple other CpG sites (each of which correlate with other CpG sites, and so on) are of particular interest as they provide a means by which we can evaluate hypothesis 1 (as described in section 1.6).

The primary application of correlation networks in our research is the identification of pathways which are connected by strong correlations in DNA methylation. Chapter 7 contains our work regarding this topic.

1.4 Cohorts

A cohort is typically described as a group of subjects with some common factor that allows them to be grouped together in some meaningful way. For example, a cohort that we might use in an EWAS could be based on age or ethnicity. Appendix B details the cohorts used for original research described in this thesis. Unless otherwise stated, they have the following characteristics:

- Same sequencing/data acquisition platform (e.g. WGBS sequencing, Illumina EPIC array)
- A defined age range
- A sufficient number of subjects such that meaningful insights can be derived from the data

In practical terms, cohorts used for studies in this thesis will generally be a collection of samples of the same tissue type from different individuals. For example, the CHDS cohort described in Appendix B consists of EPIC array data for 120 different blood samples. Cohorts can be broken down into sub-cohorts relevant to a specific purpose, such as only taking samples from individuals of a particular age.

1.5 Study Rationale

Though there are established links between DNA methylation and gene expression (e.g. Anastasiadi et al. 2018), and a vast number of biochemical pathways have been identified, the study of correlation between DNA methylation intensities at different CpG sites appears to have been largely ignored. Conventionally, epigenome-wide association studies (EWAS) look at the methylation intensities directly (as a derived beta or M value) as opposed to the correlations between them (examples: Wang et al. 2015, Sala et al. 2020). Correlation in the mathematical sense occurs when there is an association between two variables; correlation of DNA methylation between two different CpG sites would suggest the existence of an association that may be grounded in other biological processes. The biological interpretation of these correlations is discussed in section 1.3.3.

Studies in this thesis will ultimately attempt to find epigenetic trends by looking at DNA methylation derived from microarray analysis of whole blood samples from living humans. Whole blood DNA methylation data has been used in the past for a multitude of EWAS (e.g. Osborne et al. 2020; Gerring et al. 2018). The heterogeneous nature of blood can lead to some difficulties due to the cell composition effects on DNA methylation, but epigenetic regulation of processes that are not cell-mediated are typically interpretable in blood DNA methylation data (Houseman, 2015). It follows that the underlying associations of processes that are cell-mediated would still be present, if obscured by 'noise'; the effect of this noise on correlations in our data are currently unknown. If it can be shown that genes within the same pathway tend to have correlating methylation at their CpG sites, then this may lead to the discovery of new pathways as unforeseen correlations could indicate that a pair of genes comprise part of an unknown network. As such, the study may

shed some light on undiscovered regulatory pathways while also verifying pathways identified in other studies.

A number of the studies presented in this thesis are undertaken with the intention of supporting accurate evaluation of correlations in other studies. This includes data-based experiments for the following:

- Comparing the effects of microarray normalisation on correlations
- Identifying the ideal measure of correlation

This will help produce a protocol which will promote the identification of correlations which are grounded in underlying biological phenomena. Such a protocol will facilitate a more robust investigation into the hypotheses presented in this thesis and may be useful for future researchers who intend to compare correlations in DNA methylation intensity, rather than comparing the data directly as is usually the case in EWAS. Some recommendations for future work can be found in section 8.5.

1.6 Research Aims

Hypothesis 1: CpG sites on different genes which are part of the same pathway will strongly correlate in methylation intensity

The downregulation of a metabolic pathway may manifest due to an increase in DNA methylation of CpG sites on genes within that pathway, and vice versa. We can validate this hypothesis by finding stronger-than-average correlations in methylation intensity at these sites.

Hypothesis 2: CpG sites that are located closer together will tend to correlate more strongly

A topologically-associating domain (TAD) is a region in which DNA sequences physically interact with each other more frequently than with sequences outside this domain (Pombo and Dillon, 2015). It is hypothesised that CpGs on TADs, which would tend to be closer together, are more likely to have a similar methylation intensity. We can validate this hypothesis by comparing methylation intensity with CpG distance within each chromosome.

Hypothesis 3: Strong correlations in microarray-derived methylation intensities will tend to be consistent regardless of which normalisation method was used

DNA methylation microarray normalisation, discussed in sections 2.1.2 and 4.1, is a step which can improve the calculation of methylation intensities. While selection of normalisation method will certainly impact the calculated methylation intensities, it is hypothesised that strong correlations will be consistent as the underlying trends will still be present. This can be validated by comparing the strong correlations identified in a set of methylation intensities as calculated by a number of different normalisation methods.

Chapter 2

Methods

2.1 Methodological review

2.1.1 Analysis techniques

There are multiple ways of determining the methylation trends of a particular CpG site within a genome. Two of the most-used modern methods of profiling DNA methylation, based on search results for NCBI GEO (Edgar et al. 2002) are high-throughput sequencing techniques such as whole-genome bisulfite sequencing (WGBS), and the use of microarrays such as Illumina's MethylationEpic BeadChip (which will be referred to as the EPIC array in this thesis), with high-throughput techniques being used in roughly three times as many studies as arrays as of late 2020.

Sequencing techniques such as WGBS (Frommer et al. 1992) are capable of producing a full methylome of 28 million CpG sites in the case of the human genome. DNA methylation arrays provide a quick and relatively simple means of assessing the methylation of a vast number of CpG sites within a genome, though the current state-of-the-art EPIC array only looks at about 850,000 CpG sites at single-nucleotide resolution, or roughly 3% of the total number of CpG sites in the human genome (Pidsley et al. 2016). EPIC arrays are reasonably new and more data is available for the earlier Illumina Infinium HumanMethylation450 array, which has about 450,000 CpG sites (Dedeurwaerder et al. 2014). Previous studies have compared the two arrays and shown that there is significant correlation in the overlapping sites for whole blood samples, suggesting that both can be used in the same study (Solomon et al. 2018), and other studies have compared array data with WGBS data and found that derived results are also highly correlating (Pidsley et al. 2016). State-of-the-art nanopore sequencing technology has recently been used for identification of DNA methylation state, correlating greatly with results from bisulfite sequencing (Ni et al. 2019) though datasets such as this are few in number compared to datasets produced via other methods, and the relative infancy of this approach means that few software tools exist to work with nanopore-derived methylation data.

2.1.2 An overview of DNA methylation microarrays

Microarray technology has been used to assess DNA methylation for some time (Deatherage et al. 2009). The general concept of a microarray (referred to as simply 'array' throughout the majority of this thesis) is that oligonucleotides are fixed to a chip (forming 'probes'), and this chip is subjected to a DNA sample, from

which complementary strands attach to the oligonucleotides and result in a physical or chemical change that can be detected through some instrument (Illumina Inc., 2020a). In the case of Illumina's DNA methylation (Infinium) arrays, which are the source of all array-based data in this thesis, two site-specific fluorescent probes are produced for each CpG site - one to detect a methylated locus, and the other to detect an unmethylated locus. Prior to application to the chip, the DNA sample is treated with bisulfite, which converts unmethylated cytosine to uracil while leaving methylated cytosine unchanged (Clark et al. 2006). After the bisulfite-treated DNA sample has reacted with the probes on the chip, assessment of the ratio of fluorescent signals (via a laser-based imaging instrument) between the methylated and unmethylated probes can then be used to establish a methylation intensity for each CpG site (Illumina Inc., 2020b).

The vast majority of human DNA methylation array data (available on NCBI GEO) has been generated using one of three platforms produced by Illumina Inc.:

- MethylationEPIC BeadChip (referred to as the EPIC array in this thesis) - the most recent offering, with 863,904 CpG Sites (Pidsley et al. 2016)
- HumanMethylation450 BeadChip (referred to as 450k) - 482,421 CpG sites (Bibikova et al. 2011)
- HumanMethylation27 BeadChip (referred to as 27k) - 27,578 CpG sites (Bibikova et al. 2009)

An important technical consideration for EPIC and 450k arrays is that two different types of probes are employed - these are aptly named type I and type II probes. Type I probes are based on the technology used for the 27k array, and use a single colour with two different probes to generate methylated and unmethylated measurements. Conversely, type II probes use a single probe with two different colours to obtain these measurements (Wu et al. 2014). This multi-probe design was first used in the 450k array, but was carried over to the EPIC array (Pidsley et al. 2016). There can be substantial differences in methylation intensities detected by the different probe chemistries - the type II probes, while being able to be more densely packed onto the chip surface, tend to be less accurate than the type I probes (Dedeurwaerder et al. 2011).

Another important consideration is that of background fluorescence (Ritchie et al. 2007). This is a source of error which can be caused by biological phenomena such as non-specific binding of labelled sample to the surface of the array, technical issues such as optical noise from the sensor, or other factors. Correction of signal due to background fluorescence is referred to as background correction and is implemented in several of the methods discussed in section 4.1.

2.1.3 Literature review: Osborne et al. (2020) - Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort

A study by Osborne et al. (2020) investigated the effects of cannabis exposure on genome-wide DNA methylation of whole blood samples taken from the Christchurch Health and Development Study (CHDS) cohort. This study used the EPIC array to assess DNA methylation in a sub-cohort of 120 individuals, all approximately 28 years of age. While correlations between CpG methylation intensities weren't investigated specifically, the methods of analysing methylation intensity are applicable to the study being proposed.

Osborne et al. used beta values derived from the array data in an EWAS, with the intention of identifying the most differentially-methylated CpG sites for cannabis (without tobacco) users vs. controls, and for cannabis (with tobacco) users vs. controls. Beta values were calculated with the minfi package in R (Aryee et al. 2014). Techniques used for calculation of these beta values can be used in a similar manner for studies within this thesis, though some adaptations may have to be made depending on the format of available datasets.

Prior to conversion to beta values, the raw data was normalised using the NOOB procedure as implemented in minfi. NOOB, or Normal-exponential using out-of-band probes is an approach which accounts for technical variation in the background fluorescence signal produced during DNA methylation array data acquisition (Triche et al. 2013). Justification for using NOOB over other normalisation techniques was not given by the authors, nor were any other normalisation techniques used in the study as a comparison. The same data was used in another study, where the NOOB method was found to be provide the best results of all tested methods (Noble, 2021). In both cases, use of this normalisation technique yielded biologically relevant results so it should be considered as a possible technique of choice for preliminary studies and investigated further in later analyses.

A number of CpG sites were intentionally excluded from the analysis, including sex chromosomes, failed probes, sites with adjacent single-nucleotide variations that were deemed "potentially problematic" and all CpGs that did not map to a unique position on the genome. The result was that only 700,296 CpG sites were available for analysis, or only about 81% of the total number of CpG sites available on the array. This was justifiable, considering that the study was looking for individual CpG sites. A study taking a more statistical approach to the epigenome as a whole, however, would be better off retaining all CpG sites - the significant reduction in data available for analysis may have weakened potential associations or even removed them completely.

The study did not justify why whole blood was used and how blood DNA methylation relates to the methylation status of DNA contained within the brain, despite the primary findings of the study involving genes with reported roles in brain function. Houseman et al. (2015) suggest that there is evidence that whole blood methylation can be indicative of non-cell-mediated processes in other tissues, so while samples taken directly from the tissue of interest would provide a greater amount of information specific to cells in that tissue, whole blood can still provide insights into epigenetic regulation of a number of biochemical pathways in the body. Justification for taking the same approach in studies in this thesis was given in section 1.5.

In general, the study provides an excellent basis for methods to use for analysis of DNA methylation data sourced from EPIC arrays (also applicable to other Illumina arrays). The CHDS cohort is a promising dataset for correlation studies, and has been made available for use in our research.

2.1.4 Interpreting DNA methylation

As alluded to in chapter 1, DNA methylation is an imperfect process requiring a great deal of maintenance. This clearly leads to some variability in methylation of specific CpG sites within a given tissue type of an individual. Epigenetic studies use the concept of 'methylation intensity', which can be thought of as the expected level of methylation at a given CpG site. A high methylation intensity suggests that a site tends to

be more likely to methylated, whereas a low methylation intensity suggests the opposite. Techniques used in this thesis rely on the premise that technologies can provide a good estimate of methylation intensity for a given CpG site, based on some quantity of input data.

Two common representations of DNA methylation level are beta value and M-value (Du et al. 2010). The beta value is described as the proportion of methylated reads at a CpG site and can be calculated from both WGBS and array datasets. Beta values are always between 0.0 and 1.0, with a higher beta value suggesting a greater methylation intensity. The M-value is the log2 ratio of methylated reads to unmethylated reads, with an M-value of zero being equivalent to a beta value of 0.5, a positive M-value suggesting a beta value greater than 0.5 (with M = 2 equal to a beta value of 0.8) and a negative M-value suggesting a beta value less than 0.5 (with M=-2 equal to a beta value of 0.2).

Studies in this thesis use the beta value representation. This is because they are a linear variable confined to the interval [0.0, 1.0] while M-values are non-linear and do not have defined boundaries. The linearity of beta values makes statistical comparison easier and calculation of correlations won't be skewed immensely by unbounded values, as would be the case if M-values were used. To obtain the required beta values, Illumina array data will be processed using the minfi R package, available via Bioconductor (Aryee et al. 2014). Minfi and the means by which we use it is discussed further in sections 2.1.7 and 2.2.1. Selection of a normalisation method is an option provided by this package - these are discussed in chapter 4.

2.1.5 Analysis of correlation within each chromosome

Correlation is difficult to calculate for large datasets as the size of the correlation matrix scales quadratically with the number of features (in this case CpG sites) and the amount of computational effort required increases similarly, dependent on which coefficient of correlation is being calculated. Given that there are about 28 million CpG sites in total in the human genome and WGBS datasets would see all or most of these (Frommer et al. 1992), it would take a prohibitive amount of computational effort to calculate correlation between every CpG site in the human genome as over 784 trillion pairs would need to be calculated, resulting in about 6.27 petabytes of data (assuming 64-bit floating point representation). Even using all 865000 or so available from the EPIC array would result in a matrix with about 7.48×10^{11} elements which would use up almost 6 terabytes under the same assumption. While theoretically possible, this would certainly require significant high-performance computing resources to achieve in any reasonable period of time.

To simplify the process of computing and comparing correlations, we can consider only the correlation of CpG sites within individual chromosomes. A study by Ziller et al. (2014) undertook an analysis of 42 whole genome bisulfite sequencing (WGBS) datasets across 30 different human cell and tissue types. Their results suggested that only a fraction of the methylome (21.8%) changed as part of coordinated regulatory networks, and that 70-80% of the sequencing reads across the datasets provided little or no relevant information regarding CpG methylation. The study suggests that the 'dynamic' 21.8% of CpG sites tend to be co-localised with regulatory elements such as enhancers and transcription binding sites, supporting the work of Thévenin et al. (2014). This supports the idea to limit calculation of correlations to CpGs co-located on the same chromosome - we would still expect to find some significant correlations, assuming hypothesis 1 was correct. We must also define a threshold for a 'strong' correlation - this, as well as a more-comprehensive discussion on matters relating to correlation, can be found in section 1.3 and within the studies of chapter 5.

2.1.6 Use of existing DNA methylation datasets

Online databases such as NCBI GEO (Edgar et al. 2002) and academic institutions can provide DNA methylation data for research use. By using data from previous experiments, there is no need to produce new data – this vastly simplifies the process of obtaining sufficient data to test hypotheses and mitigates the cost of WGBS or array profiling.

WGBS and Illumina's arrays only look at cytosine methylation. The overwhelming majority of mammalian DNA methylation research limit their scope to this, though recent research suggests that modification of adenine to N6-methyladenine also occurs to a very limited extent in mammals (Wu et al. 2016). Datasets for adenine methylation are extremely limited so this study will focus entirely on cytosine methylation.

For the purposes of our research, DNA methylation data will be sourced from existing datasets, including those directly available to the University of Canterbury and collaborators, as well as publicly-available methylation databases which include information on the age of the characterised individual, such as those described by Komaki et. al (2018) and Li et. al (2018).

Appendix B describes the datasets (cohorts) used for studies in this thesis.

2.1.7 Software tools

The R programming language (Ihaka and Gentleman, 1996) is a common language used for statistical computing. Minfi (Aryee et al. 2014) is a module for R that provides functionality for processing raw data from Illumina methylation arrays, and has been used in multiple recent EWAS (e.g. Osborne et al. 2020; Noble 2021).

The Python programming language (van Rossum and Drake Jr., 1995) is a common language used for programming in a multitude of different fields. Python 3 is the most recent major version and is referred to throughout this thesis more generally as 'Python'; the specific minor version of Python used for studies in this thesis is detailed in Appendix A. Subjectively, Python presents a great deal of advantages over R in terms of ease-of-programming, flexibility and available libraries, so will be used preferentially over R wherever possible. Several open-source Python libraries that provide key functionality for our research have been identified:

- numpy (van der Walt et al. 2011) - numerical computation
- scipy (Virtanen et al. 2020) - general scientific computing
- scikit-learn (Pedregosa et al. 2011) - general scientific computing
- pandas (McKinney, 2010) - data science library, interoperable with numpy and scipy
- matplotlib (Hunter, 2007) - 2D graphics
- networkx (Hagberg et al. 2018) - network analysis, interoperable with matplotlib

Versions of all software tools used are recorded in appendix A.

2.2 General methods of calculating correlations in array-derived DNA methylation data

2.2.1 Array normalisation, preprocessing and beta value extraction

Raw array data, in IDAT format, is processed using the minfi R module (Aryee et al. 2014). This data is provided as a red/green colour set - the technical aspects of this output were reviewed for the HumanMethylation450 platform (Dedeurwaerder et al. 2014), which was the technological basis for the MethylationEPIC array used by the CHDS cohort (among others). This is accomplished with several steps:

1. The sample sheet for the raw data is read in. This is only an option for datasets which come with a sample sheet (or for which one has been generated). For cohorts without a sample sheet, minfi is able to find all two-colour IDAT files in a specific directory.

2. The set of experimental data, in the form of the two-colour IDAT files, is read in.

3. Preprocessing is performed on the two-colour IDAT files. Depending on the function used for this, normalisation can be included in this step. The normalisation methods made available by minfi include all of those discussed in section 4.1. It is also possible to generate a set of preprocessed data that is not normalised - this is referred to as 'raw' preprocessing.

4. Preprocessed data is annotated using the appropriate manifest. At this point, SNP information can optionally be added if required.

5. Methylation intensity for all CpG sites available on the array is extracted as a beta matrix. It is also possible to extract other measures of methylation intensity, such as M-values.

Further analysis is conducted using Python, so after producing a set of beta values for any given normalisation type, they are exported into CSV format for portability.

Unless otherwise stated, research in this thesis always uses System 1 and R configuration 1 (from Appendix A) for steps described in this subsection.

2.2.2 CpG subset selection and calculation of correlation

Beta values for all CpG sites on the array, for all samples, are referred to as the 'beta matrix' and are contained within a single CSV file. Much of the research presented in this thesis looks at correlations within specific chromosomes, so a script was written in Python to split this file up based on chromosome. A similar approach can be taken to arbitrarily subset the overall beta file; for example, all CpG sites on genes associated with the same pathway can be extracted into its own file, if required. This can also yield a CSV file of beta values for all autosomal chromosomes (i.e. excluding sex chromosomes). Regardless of which CSV file is used, correlations are generated by reading it in to Python using the *pandas* module (McKinney, 2010) as a dataframe and using the appropriate function, such as the `corr` method of the *dataframe* class. The output of this is referred to as the 'beta correlation matrix', abbreviated as BCM. This can be used directly by other Python functions, but if it is likely to be required for multiple studies, the matrix is serialised using Python's inbuilt *pickle* module (usually via a function in the *pandas* module) and saved to disk for later use to reduce future computational load and time requirements.

When using the `corr` method to calculate correlations, a measure of correlation coefficient must be specified. Spearman's rank correlation is used in this thesis unless specified otherwise. Reasons for this are discussed in section 1.3.2 and studies within chapter 5.

2.2.3 Identification of a CpG site's associated gene

The genes associated with CpG sites are included in the array manifest provided by Illumina. This manifest is available on the Illumina website on their product support page. We can use the manifest directly to obtain the name of the associated gene and use these as required, though not all CpG sites will necessarily have an annotated gene, and the manifest may not be regularly updated to make use of the newest research. In that case, we can consult an external database such as GENCODE (Harrow et al. 2012) or an online tool such as the UCSC genome browser (Karolchick et al. 2003) may be used to manually identify which gene a CpG site is associated with.

2.2.4 Statistical analysis of beta values and beta correlation matrices

For the purposes of comparison and evaluation, we can consider the following common metrics for both beta values and their associated correlation matrices:

- Average/mean: a measure of central tendency of a dataset, generally influenced by all values equally unless a weighted average is taken
- Median: an alternative measure of central tendency, based on the value of the middle number of the dataset when ranked in order (or the mean of the middle two, for an even-numbered dataset)
- Standard deviation: a measure of the amount of variation within a set of variables
- Variance: an alternative measure of variation within a set of variables, equal to the standard deviation squared
- Skew/skewness: a measure of asymmetry of the distribution of a dataset, with a positive skew indicating that the primary 'tail' of a distribution is on the right, and a negative skew indicating that the primary tail is on the left, for a unimodal distribution. We used the unbiased skew measure in this thesis.
- Excess kurtosis: a measure of the 'tailedness' of the distribution of a dataset. An excess kurtosis of zero generally suggests a normal distribution (e.g. Gaussian). A positive excess kurtosis suggests the distribution has fatter tails and a more densely-packed centre, and a negative excess kurtosis suggests the distribution has thinner tails and a less densely packed centre. We use the unbiased excess kurtosis in this thesis.

Additionally, we define a few more metrics that aren't in common use:

- Ratio of positive to negative correlations: a measure of the tendency of a method and dataset to generate positive correlations, rather than negative ones (with a ratio of greater than 1.0 suggesting more positive correlations, and vice versa). Self-correlations (i.e. the diagonal of the correlation matrix) are excluded from this calculation. We use this as a basic gauge of the spread of values within a correlation matrix, with a ratio closer to 1.0 implies a more even distribution of positive and negative correlation values.

- Consistent and unique strong correlations for a proportion: a measure of similarity between two or more subsets of strong correlations, where subsets could be taken from matrices generated from data produced by a different normalisation type, cohort, etc. Values are sorted and the top and bottom X% are taken for comparison, where X is the selected proportion. The number of consistent correlations is the number of pairs that correlate strongly in all subsets. The number of unique correlations, per subset, is the number of pairs that are only correlate strongly in that subset.

- Offsets in mean and variance/standard deviation within multiple datasets: for a set of n subsets of data, with an overall mean of μ and overall standard deviation of σ, we take a subset's offset in mean as μ subtracted from that subset's mean, and a subset's offset in standard deviation as σ subtracted from that subset's standard deviation (alternatively, a subset's offset in variance is σ^2 subtracted from that subset's variance). These offsets can be compared to determine the relative effects of whatever the difference is between the subsets, such as comparing effects due to normalisation type.

There are a number of options for evaluating statistical properties. Several Python libraries, such as *pandas*, *numpy* and *statsmodels* (among others) provide excellent statistical capabilities and integrate nicely into the larger Python-based data pipeline that we are developing for this research.

2.2.5 Selection of strong correlations

Our two options for selection of strong correlations (both discussed in more detail later in this section) are threshold-based, where we take all correlations with a magnitude greater than some arbitrary threshold; and proportion-based, where we take some percentage of the strongest positive and negative correlations. In either case, we must exclude the following two subsets within the correlation matrix prior to selecting strong correlations:

1. The correlation matrix diagonal, which are correlation scores of one CpG with itself (trivially 1.0 in all cases).

2. Either the upper or lower triangle of the matrix, as one is the duplicate of the other.

Either of these exclusions can be performed easily with functions built-in to the Python libraries we are using for our research, or implemented algorithmically into the selection process.

2.2.5.1 Threshold-based selection

We can define a strong correlation as a correlation coefficient with magnitude greater than or equal to some number between 0 and 1.0. A script was written to pull out values of at least this strength from a beta correlation matrix with the two associated CpG sites. These correlations are saved to a file for later use and/or used immediately by other scripts if required.

As the number of a chromosome's correlations above this threshold will vary based on factors such as normalisation type and cohort, it is more difficult to use threshold selection for direct comparison - when comparing the strongest correlations for a given chromosome (e.g. for comparing normalisation types or cohorts), a better alternative would be to use a proportion-based method (described in section 2.2.5.2) as this will ensure the same number of strong correlations are available for comparison.

For the purposes of this thesis, the threshold for a strong correlation is defined as 0.7 unless otherwise stated.

2.2.5.2 Proportion-based selection

An alternative approach to selecting strong correlations is to select a subset of correlations which have values closest to the extremes; i.e. some number of correlation coefficents which are closest to 1.0 (for positive) or -1.0 (for negative). This number can be arbitrarily selected, e.g. the 1000 values closest to 1.0, and the 1000 values closest to -1.0, or based on the size of the correlation matrix, e.g. the top and bottom 10% of correlation coefficients if they were arranged in order from most positive to most negative.

The key advantage of a proportion-based threshold is that we essentially specify how many of the strong correlations we want. This allows for easier comparison between different methods. The technical cost of a proportion-based selection is significantly higher, however, as the sorting procedure becomes more computationally-intensive as the size of the dataset becomes larger - by default, Python's sorting algorithm is *Timsort*, which has a worst-case complexity of $\mathcal{O}(n\,log(n))$ (Auger et al. 2018[1]). Due to this, a proportional-based threshold might not be computationally-feasible for large datasets. Another concern is that there may be an insufficient number of total positive or negative correlations for a given proportional threshold; i.e. if we wanted to take the 10% strongest positive and negative correlations, but the ratio of positive to negative correlations was such that the 10% most negative correlations includes some values greater than 0.0, we may run into some difficulty. Therefore, we should also keep in mind this ratio when using a proportional threshold.

2.2.6 Correlation network analysis

Correlation networks are defined in section 1.3.4.1. To conduct our network analysis, we use the *networkx* module for Python (Hagberg et al. 2008). This module allows us to create undirected graphs for network analysis without having to change programming language.

Networks are constructed by first adding all CpGs with positive correlations as vertices. Where necessary (such as for analysis of genes on specific pathways), genes that these CpG sites are associated with can be used as vertices instead - they simply inherit the connections of their constituent CpG sites; consequently, if two correlating CpG sites are associated with the same gene, it is possible for the consolidated gene to be an orphan vertex. Either way, edges between the positively correlating CpGs/genes are added next. The overall graph is then updated with negative connections in the same manner, using the positive correlation graph as a basis.

To aid in visually representing the correlation network, the following colours are defined:

- A positive correlation is represented by a green line
- A negative correlation is represented by a red line
- In the event that two vertices derived from genes share at least one positively-correlating pair of CpGs and one negatively-correlating pair of CpGs, the line between these genes is a shade of grey or black

[1] arXiv preprint

'Green' and 'Red' are not strictly defined but the differences should be apparent in all graphs shown in this thesis. Other graph elements remain undefined and are displayed based on how they can best convey information.

Graph statistics, such as numbers of nodes and edges, can be extracted from graphs using *networkx* functions such as the `number_of_nodes` method of the Graph class. Complex operations may be available in the *networkx* library but are implemented manually where required.

2.3 Automated acquisition of gene and pathway information

Biological pathways arise from the interplay between different genes; as such, we have a very large and complex set of genes and pathways to work with. Estimates of the number of protein-coding genes within the human genome have varied greatly over the years, with recent estimates placing the number of protein-coding genes at about 21,000, non-coding genes at about 22,000, and total number of transcripts to be on the order of 323,000 (Pertea et al. 2018). Manually going through all of these genes would be quite an endeavour, so an automated approach is employed.

Several relevant online databases have been identified as having potential use in an automated search context:

- GENCODE (Harrow et al. 2012) - gene features in the human genome
- UCSC genome browser database (Karolchik et al. 2003) - annotated genome sequence information
- Reactome (Fabregat et al. 2018) - biological process information as a network of molecular transformations

Our knowledge of human metabolism is far from complete, and online databases will reflect not only these gaps in our knowledge, but also any known genes and pathways that haven't been added to them (intentionally or otherwise). This is something we must keep in mind when discussing results from studies that make use of automatically-searched data.

2.3.1 Selection of chromosomes based on HPC limitations

As discussed in section 2.1.5, correlation matrices require substantial computational processing power to generate. Much of the work in this thesis is undertaken on a virtual machine residing in the University of Canterbury's Research Compute Cluster. System specifications for this virtual machine are detailed as System 2 in Appendix A. As HPC resource is not infinite, we have a limited amount of computational ability for our studies.

Generally, use of the uncorrelated beta matrices will not present an issue as the largest one (comprising all chromosomes) is a mere 1.8GB for our preferred test cohort (CHDS, described in Appendix B). This matrix consists of roughly 860,000 beta values for 120 individuals and can easily be manipulated and analysed with the 128GB of available RAM on our virtual machine. Section 2.1.5 described how this would result in a correlation matrix about 6TB in size - generating this would not be feasible on our current resource. We may run into problems working with larger chromosomes when correlation matrices are involved, especially if we are performing operations with memory requirements which scale with the size of the dataset.

Where the analysis of more than one chromosome is required (e.g. the studies in parts II and III), the best approach is to start with those that generate the smallest correlation matrices and work our way upwards in size from there. In the context of DNA methylation microarrays, correlation matrix size is quadratically-proportional to the number of CpG sites probed at each chromosome. The number of probes for each chromosome, arranged in order of ascending size, is shown in table 2.1.

The computational requirements for a specific task will vary; a simple search of a correlation matrix will not require much more than what was needed to generate it, but operations involving sorting or large-scale statistical analysis may be significantly (orders of magnitude) more computationally- and/or memory-intensive. As such, there is no strict rule around how many chromosomes should be used for a study, other than the general 'as long as it fits on the HPC, it'll be fine'.

Chromosome	**Number of Probes**
y	537
21	10300
18	14899
22	18367
x	19090
13	21040
20	22960
9	26167
15	28741
14	29550
4	36771
16	37939
8	38452
19	38550
10	42126
17	44435
12	44623
5	44720
7	47560
11	48894
3	48896
6	54401
2	64828
1	82013

Table 2.1: Number of probes for each chromosome, for the EPIC array.

Part II

Technical Studies

Chapter 3

A preliminary assessment of correlations in DNA methylation

3.1 Premise

In this chapter, we use published DNA methylation data to develop a computational pipeline through which we are able to statistically analyse methylation intensities (in the form of beta values) at different CpG sites, and correlations between them, using available data for chromosome 21. Specifically, we investigate:

- The statistical distribution of beta values for CpG sites on chromosome 21
- The statistical distribution of beta matrix correlations on chromosome 21
- Analysis of the strongest beta matrix correlations on chromosome 21
- Potential interaction networks that we can infer from beta matrix correlations on chromosome 21

In order to develop the tools, it is necessary to make some assumptions. For example, we assume that findings from a study of chromosome 21 will also be relevant to other chromosomes and that the choice of test cohort does not introduce significant sources of error, such as selection bias. This broad approach does not test any particular hypotheses but rather focuses on development of methods suitable for further studies. We therefore acknowledge these limitations and take them into consideration when deriving insights from this preliminary study.

The CHDS cohort (see appendix A) is very well-documented and has been used in other epigenetic studies (e.g. Osborne et al. 2020), so serves as a suitable test cohort for the purposes of method development and proof-of-concept.

3.2 Study: Beta values for Chromosome 21

Prior to studying correlations, an initial investigation into the statistical distribution of beta values for CpG sites on chromosome 21 was undertaken. This chromosome was selected as it was the autosomal chromosome with the fewest number of CpG probes in the EPIC array (10300) - this would be beneficial for a future

correlation study (conducted in section 3.3), as would understanding the distribution of beta values so we can make informed choices on how to approach and interpret the resulting correlation matrix.

Two sets of beta values are looked at - those produced via the NOOB method of normalisation, and those produced without prior normalisation ('raw'). NOOB was selected on the basis that it has been shown to perform better than other normalisation methods in some circumstances (Noble, 2021). Raw betas were selected for comparison.

3.2.1 Methods

Beta values were computed using the protocol described in section 2.2. Statistical analyses used the Python pandas library.

This study used System 1 and Python configuration 1 as described in Appendix A.

3.2.2 Results and Discussion

3.2.2.1 Statistical distribution of means

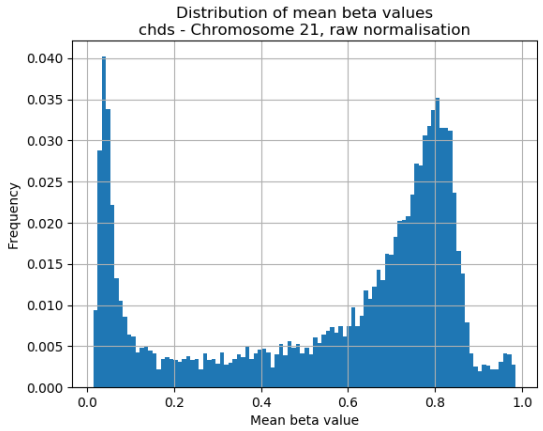

Figure 3.1: Histogram of mean beta values - CHDS cohort, Chromosome 21, no normalisation (raw)

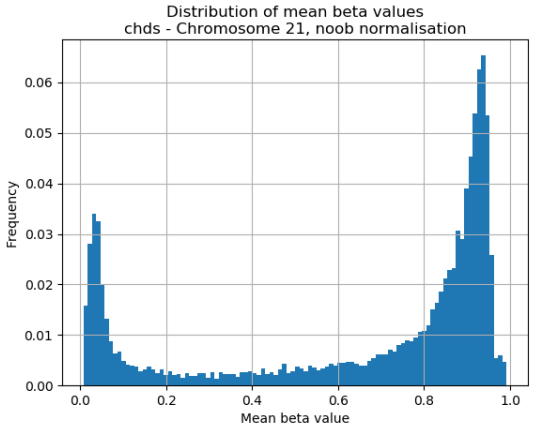

Figure 3.2: Histogram of mean beta values - CHDS cohort, Chromosome 21, NOOB normalisation. Please note that the scale of the y-axis is different between figures 3.1 and 3.2.

Figures 3.1 and 3.2 show that both NOOB-normalised and unnormalised raw beta values from samples within this cohort tend to be highly-biased towards the extremes of 0.0 and 1.0. The NOOB-normalised betas seem to follow a J-shaped distribution with a prominent peak at the high-beta end, compared with the U-shaped distribution in the unnormalised betas which has a much broader and flatter high-beta peak. The region between the two peaks on both histograms is significantly less-populated, suggesting that samples tend to be more generally strongly methylated or unmethylated rather than expressing an intermediate level of methylation at a particular locus. This makes sense biologically as genes would ideally be 'switched off' in places where they are not needed and one mechanism by which this can be accomplished is by methylating

specific parts of that gene (for reasons discussed in section 1.1). Genes that produce more-commonly required products would tend to be unmethylated for the same reason. The central 'flat part' may represent loci that are weakly switched on or off depending on the dominant process within their respective tissue type; whole blood methylation can be indicative of non-cell-mediated processes in other tissues, as described in section 2.1.3.

The overlaps within the two peak regions of each normalisation method were also calculated to observe how selection of method impacts the beta values in these regions. Rather than defining thresholds, the CpG sites corresponding to the lowest 2000 and highest 6000 beta values were compared for each normalisation type. These numbers were chosen based on visual inspection of Figures 3.1 and 3.2, and represent about 20% and 60% of all probed betas for chromosome 21:

- Low beta overlap: 1954/2000 (97.7%)
- High beta overlap: 5871/6000 (97.9%)

The high beta and low beta overlaps are both very high, with only ~2-3% difference between the raw and NOOB-normalised beta values. This is good, as it means that normalisation type doesn't greatly alter which CpG sites have the highest and lower beta values. It doesn't tell us if correlations will be consistent, however; we investigate this in a later study.

3.2.2.2 Statistical distribution of standard deviations

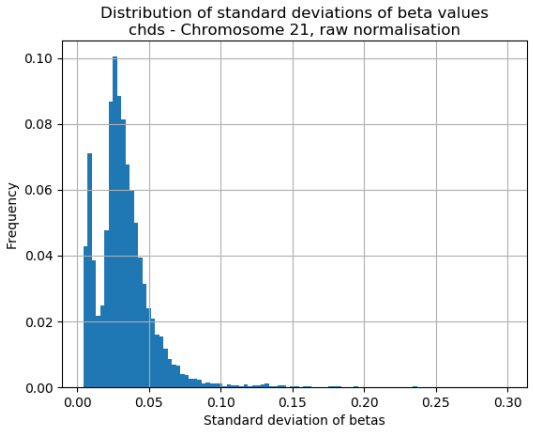

Figure 3.3: Histogram of standard deviations of beta values - CHDS cohort, Chromosome 21, no normalisation (raw)

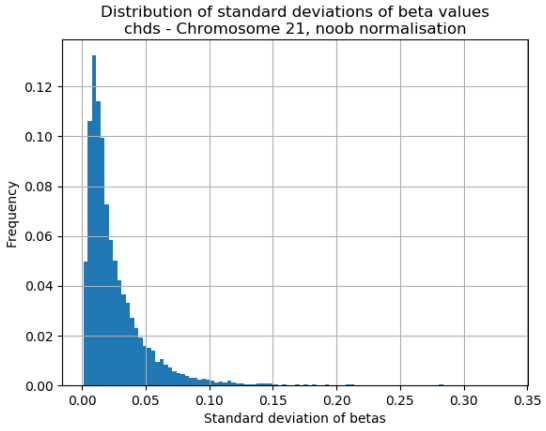

Figure 3.4: Histogram of standard deviations of beta values - CHDS cohort, Chromosome 21, NOOB normalisation

Figures 3.3 and 3.4 give an indication of how much the beta values can vary at each CpG site. We see that the raw betas have a notable cluster of sites with a comparatively-low standard deviation, preceding a large peak that can also be found in the NOOB-normalised distribution (albeit shifted slightly). Figure 3.4 suggests that NOOB normalisation decreases the standard deviation (and therefore variance) of some number of CpG sites, based on the peak in Figure 3.4 being shifted to the left. This gives us slightly more certainty over the accuracy of our measure of methylation intensity at these CpG sites as the true value

(or population mean) is more likely to sit within a narrower range of values - in statistical terms, a lower standard deviation equates to a narrower confidence interval in all practical cases, and we can interpret this as a general increase in certainty of our derived beta value.

3.3 Study: Correlations on Chromosome 21

An initial investigation into the statistical distribution of correlations on chromosome 21 was undertaken to identify factors that may be relevant to correlation analyses, such as selection of thresholds for strong correlations.

The same beta values produced for the study in section 3.2 are used for this study, for the same reasons.

3.3.1 Methods

Beta values and correlation matrices were computed using the protocol described in section 2.2. We assess the distribution of correlation coefficients with a histogram. The statistical analyses in this section use a dataset comprised of the average for each correlating pair, and were conducted using the Python pandas, scikit-learn and numpy libraries.

This study used System 1 and Python configuration 1.

3.3.2 Results and Discussion

3.3.2.1 Statistical analysis of correlations

Histograms (Figures 3.5 and 3.6) showing correlation frequency was produced for both beta correlation matrices. Some statistical metrics are presented in Table 3.1.

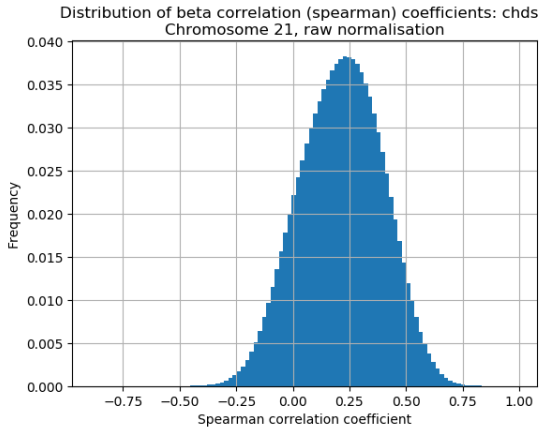

Figure 3.5: Histogram of beta correlation (Spearman) coefficients - CHDS cohort, Chromosome 21, no normalisation (raw)

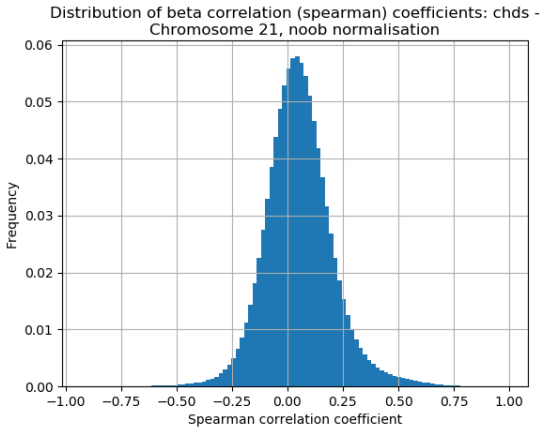

Figure 3.6: Histogram of beta correlation (Spearman) coefficients - CHDS cohort, Chromosome 21, NOOB normalisation. Please note that the scale is different between figures 3.5 and 3.6.

	raw	**NOOB**
mean	0.214	0.053
median	0.218	0.046
ratio of positive to negative correlations	6.69	1.77
standard deviation	0.183	0.154
skew	-0.103	0.344
excess kurtosis	-0.195	2.061

Table 3.1: Statistical analysis of all beta matrix correlations calculated for Chromosome 21 (CHDS cohort; n = 120), for raw (no normalisation) and NOOB normalisation.

While both appear to follow a symmetrical (though offset) distribution, the differences in distribution of correlation coefficients are very apparent in Figures 3.5 and 3.6. Both distributions have a mean and median value greater than zero, and the ratio of positive to negative correlations in both cases is greater than one - this suggests that CpG sites are more likely to correlate positively than negatively as there are more (and stronger) positive correlations than negative ones. We can see that this is much more the case for the unnormalised (raw) data in spite of the negative skew value. This is most likely due to background fluorescence which has not been accounted for - as the sensor is based on fluorescent signals, background fluorescence would likely be causing the methylation signal to be stronger than it should be, shifting the distribution of methylation intensities at each CpG site to the extremes and thus producing stronger correlations for beta values, as they are constrained to a maximum magnitude of 1.0. An alternative explanation could be technical bias, such as effects due to CpGs being on the edge of the chip, or batch effects that may only affect certain CpGs.

The increased excess kurtosis of the NOOB histogram suggests it may be easier to identify strong correlations in NOOB-normalised betas than in those without normalisation as the broader tails provide a

	raw	NOOB
Number of strong positive correlations	98644	66137
Number of strong negative correlations	661	16284
Total number of strong correlations	99305	82421

Table 3.2: Number of strong positive and negative correlations derived from beta correlation matrices, per norm type

greater range from which we can select a cut-off for these strong correlations. Arbitrarily defining a strong correlation as one of magnitude 0.7 or greater (as in section 1.3.4) is an approach that we could take (e.g. if we visually selected one based on Figure 3.6), but we may be better off taking an approach based on the distributions themselves. For example, we could take a subset of the correlations closest to 1.0 and -1.0. This is explored in section 3.3.2.2.

3.3.2.2 Strongest correlations

Table 3.2 shows the number of strong (magnitude ≥ 0.7) positive and negative correlations taken from the beta correlation matrices calculated from the unnormalised (raw) betas and normalised (NOOB) betas. About 0.09% of all correlations were classified as strong. We saw in section 3.3.2.1 that the unnormalised beta values were biased towards producing positive correlations, and by counting the number of strong correlations, we can see that it is rare for values within the unnormalised beta correlation matrix to be less than or equal to -0.7. Though not nearly as severe, a similar effect can be seen for the NOOB-normalised data, with about 80% of the strong correlations being positive. This is something that will need to be considered when selecting thresholds for later correlation studies.

Table 3.3 shows the number of overlapping strong (magnitude ≥ 0.7) positive and negative correlations taken from the beta correlation matrices calculated from the unnormalised (raw) betas and normalised (NOOB) betas.

	raw	NOOB
Overlapping positive correlations	41493/98644 (42.1%)	41493/66137 (62.7%)
Overlapping negative correlations	658/661 (99.5%)	658/16284 (4.04%)

Table 3.3: Overlap of strong positive and negative correlations derived from beta correlation matrices, per norm type (magnitude threshold for strong correlation = 0.7)

Overlap of positive correlations is inconsistent, as is the overlap in negative correlations. The percentage of overlapping negative correlations is being skewed significantly due to the difference in number of negative correlations of each method (661 versus 16284) as there simply are not enough strong negative correlations within the raw betas to make for a viable comparison. As an alternative, we can consider some number of strong positive and negative correlations rather than setting an arbitrary threshold, as this will ensure that the numbers of strong positive and negative correlations are consistent between different normalisation types. This approach is described in section 2.2.5.2. Table 3.4 shows the results if we look at the strongest 1% of positive and negative correlations, for a total of 1060900 of each.

	raw	NOOB
Positive overlap	303824 (28.6%)	
Negative overlap	341113 (32.2%)	
Positive range	[0.56937, 0.98472]	[0.42089, 0.98858]
Negative range	[-0.15896, -0.882033]	[-0.24806, -0.919633]

Table 3.4: Overlap of strong positive and negative correlations derived from beta correlation matrices, per norm type (1% strongest positive and negative correlations)

We can see a $\sim$ 30% overlap in 1% of the data, but taking the top 1% lowers the equivalent threshold of a strong negative correlation to -0.15896 for the raw beta values. This has a much lower magnitude than the threshold for a strong correlation defined as 0.7 in section 2.2.5. Further research is required to work out if we can find biologically-relevant correlations with such a low effective threshold.

3.3.2.3 Network analysis of strong correlations

To determine the value of correlation analyses in identifying biologically relevant correlations between CpG sites, a network analysis was conducted using correlations with an absolute strength of 0.7 or greater. Methods for conducting this analysis are described in section 2.2.6. For this analysis, each CpG site was represented as a node on a graph, with edges representing a strong correlation (coloured red for negative, and green for positive correlation). The overall graph for both methods is hard to resolve in a way that works in a written format, so CpG labels were removed for the purposes of visualisation in this subsection. These can be seen in Figures 3.7 and 3.8. Notably, these graphs consist of a primary core network and a large number of very small 'satellite' networks.

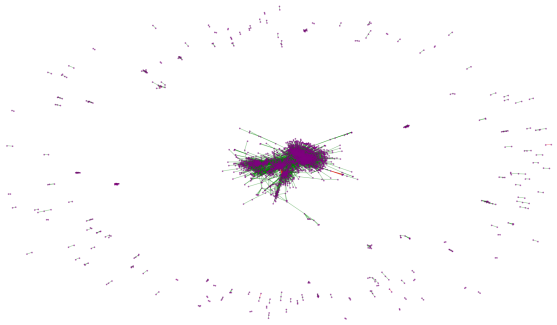

Figure 3.7: Unlabelled graph of strong beta correlations on chromosome 21 (unnormalised betas)

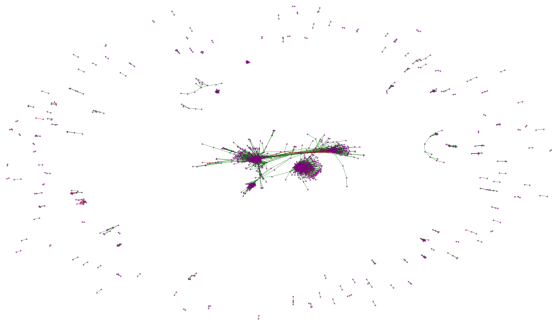

Figure 3.8: Unlabelled graph of strong beta correlations on chromosome 21 (NOOB-normalised betas)

As per section 2.2.3, we can convert the CpG vertices to their annotated gene. Using methods discussed in section 2.2.6 we can focus on a specific gene and only look at other genes connected (via correlation) within a certain number of edges, i.e. sub-networks centered on a gene of interest and looking at a defined 'neighbourhood' of correlating genes and CpG sites. As an example, Figure 3.9 shows a subnetwork of the NOOB-normalised network, focused on the KRTAP6-1 gene and its immediate neighbours. This gene was

first characterised by Rogers et al. in 2002 and is one of a number of keratin-associated proteins found on chromosome 21.

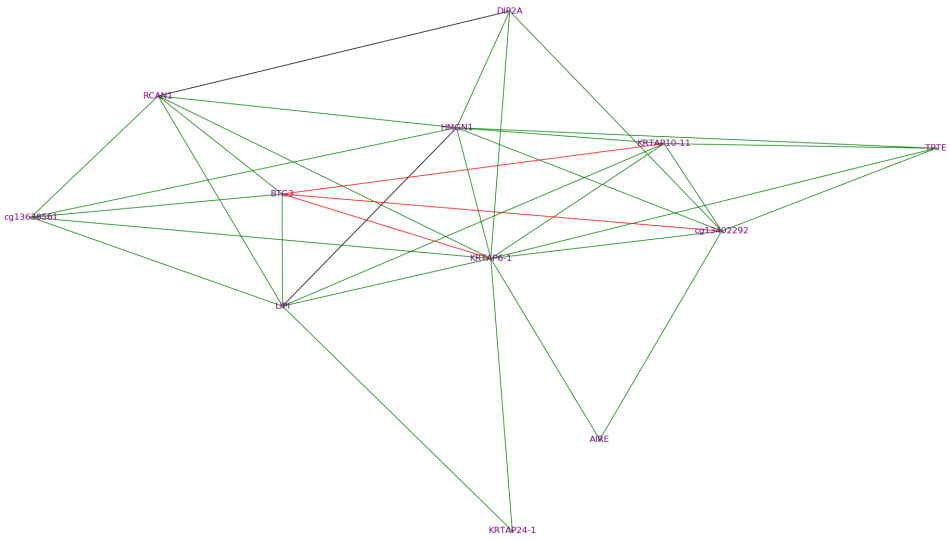

Figure 3.9: Subnetwork of keratin-associated protein 6-1 (KRTAP6-1) and its immediate highly-correlating neighbours, taken from the network derived from highly-correlating pairs within the NOOB-normalised beta correlation matrix.

From Figure 3.9 we can see that several keratin-associated proteins have CpG sites which tend to correlate positively in methylation intensity; based on this, the methylation intensity of KRTAP6-1 would appear to correlate with both KRTAP10-11 and KRTAP24-1. Additionally, the negative correlation between methylation intensity of BTG3 and both KRTAP10-11 and KRTAP6-1 is link which is not particularly intuitive, given that BTG3 (B-cell translocation gene 3) is known more for its alleged anti-proliferative properties (Winkler, 2010) rather than its action with keratin-associated proteins. The relationship between BTG3 and keratin-associated proteins has not yet been studied. However, data in this study suggests that our understanding of these genes may benefit from further research.

3.4 Concluding remarks

This preliminary assessment has investigated the potential methods and implications for DNA methylation intensity correlation analysis. The approaches described in section 2.2 have indicated that this methodology is able to detect correlations among CpG sites that may be useful for identifying biologically-relevant correlations between sites, and will serve as a basis for other studies discussed in this thesis.

Using the NOOB method of normalisation produce results that were more-easily delineated than those produced using unnormalised data when it came to assessing correlations, though a more robust comparison is required - chapter 4 investigates this further. Further research into how to determine strong correlations is also required as arbitrary setting a threshold resulted in a very imbalanced spread of negative and positive values. An alternative to this is taking some percentage of the strongest and weakest correlations. We investigated this initially in section 3.3.2.2 and take this investigation further in chapter 5.

We have shown that we can build networks from CpG sites that correlate in methylation intensity. Network analysis yielded graphs that suggested there may be some large and complex networks between correlating CpG sites, but this is something that requires significantly more research effort to investigate - chapter 7 goes into correlation network analysis in more depth. Our work is not done, however. To fully take advantage of the research potential of these networks, we will need to undertake further studies to aid in our selection of methods used for array normalisation and calculation of strong correlations. These in particular are investigated in chapters 4 and 5.

Chapter 4

An assessment of array normalisation method choice

4.1 Selected array normalisation methods

The general ideas behind DNA methylation microarrays and array normalisation were discussed in section 2.1.2. Further to that, this study has investigated five methods of array normalisation, as well as a sixth approach which used raw data, referred to in this thesis as 'raw normalisation', 'unnormalised', etc. as appropriate. A detailed description and explanation of each of the individual methods is beyond the scope of this thesis, however each of the five normalisation types is briefly described in the following subsections. The sixth 'raw' normalisation method simply uses data as presented by the array and does not require further explanation.

4.1.1 Normal-exponential using out-of-band probes (NOOB)

This method was described by Triche et al. (2013) as an extension of an earlier approach to correcting for background fluorescence (Ritchie et al. 2007). The general idea is that the background effects are estimated using a set of control probes, taken from a set of known negative controls and out-of-band type I probes. A probabilistic approach is taken to determine the expected (exponentially-transformed) signal parameter, given an observed foreground intensity. An enhancement to NOOB was described that removed the need for a reference sample (Fortin et al. 2017) - this is referred to as ssNOOB and was incorporated into minfi's implementation of NOOB.

4.1.2 Stratified quantile normalisation (Quant)

Stratified quantile normalisation was initially proposed by Touleimat and Tost (2012) to overcome the issues presented by the two different probe chemistries present on the HumanMethylation450 microarray. Quantile normalisation is a statistical technique which has been used in genomics studies for some time (Amaratunga and Cabrera, 2001) which, in short, alters statistical distributions so that they have the same statistical properties. The premise of this method involves using the signals from type I probes as 'anchors' to normalise type II probes, producing results that were considered significantly better than those generated using other normalisation methods in use at the time (Touleimat and Tost, 2012).

4.1.3 Functional normalisation (FunNorm)

Functional normalisation was proposed as an extension to Stratified Quantile normalisation. Where quantile normalisation forces the empirical marginal distributions of the samples to be same, removing all variation in the statistic, functional normalisation makes use of reference covariates (recommended as the first m=2 principal components of the control summary matrix by the method's author) when removing variation, thus reducing some of the technical effects present in quantile normalisation (Fortin et al. 2014).

4.1.4 Subset-quantile within array normalisation (SWAN)

The SWAN method was described by Maksimovic et al. (2012). Similar to the Quant method, it is a between-array normalisation technique, though it uses a random subset of probes to do the between-array normalisation. Maksimovic suggests that SWAN provides an advantage for modern Illumina arrays, as the technique accounts for differences between type I and type II probes on a single array, allowing them to be normalised together with fewer consequences that other methods.

4.1.5 Illumina's method

The minfi library has a reverse-engineered implementation of the normalisation method used in their Genome Studio tool. The original algorithm is described in the GenomeStudio Methylation Module v1.8 User Guide (Illumina Inc. 2010). This method requires a reference array to be selected manually. The documentation in minfi doesn't specify what to do in the absence of a defined reference. In the context of this thesis, this method will be used for comparative purposes but won't be considered for any serious biological study owing to its lack of documentation.

4.2 Research trends

It would be difficult to directly assess the relative popularity of the normalisation methods investigated in this study, so we've opted for a naïve comparison of citations for each method's original publication, based on PubMed and Google Scholar. The source for Illumina's method is not available on either of these platforms so is not considered here. This data was taken on 16 December 2020.

Method	PubMed Citations	Google Scholar Citations
NOOB	207	387
ssNOOB	123	204
Quant	176	370
FunNorm	226	414
SWAN	342	622

Table 4.2.0: Comparison of PubMed and Google Scholar citations for papers first describing selected normalisation methods

As we can see, Maksimovic's paper for SWAN is the most cited of these, even if we combine the number of citations for NOOB and ssNOOB. We interpret this as SWAN being the most popular method, though this doesn't necessarily imply that it is the best method.

4.3 Computational considerations

To compare the time taken for each of the five methods (and raw preprocessing), they were subjected to computational profiling. For testing, the six methods of preprocessing (five different normalisation methods and the sixth method of 'no normalisation') described in section 4.1 were conducted sequentially, each using unprocessed array data from the CHDS cohort (see appendix B), and the computational time for each method was profiled. This was run twice, with the second time running the methods in a different order, to ensure there were no 'under the hood' effects due to order. This analysis was conducted using System 1 using R Configuration 1, both described in appendix A, and all functions used default arguments except for the input *RGChannelSet* which was derived from the raw IDAT data.

Method	**Runtime 1 (s)**	**Runtime 2 (s)**	**Average Runtime (s)**
preprocessRaw	3.24	2.37	**2.81**
preprocessNoob	99.03	100.30	**99.67**
preprocessSWAN	74.40	75.67	**75.04**
preprocessFunnorm	210.51	212.01	**211.26**
preprocessQuantile	69.05	68.62	**68.84**
preprocessIllumina	14.24	13.74	**13.99**

Table 4.3.0: Computational profiling of selected normalisation methods: time taken (in seconds) to generate normalised data for Chromosome 21.

Ignoring the trivial case of no preprocessing (*preprocessRaw*), the fastest results are achieved with the Illumina-based algorithm. Conversely, functional normalisation has the highest runtime by a significant margin, at over twice the runtime for the next-slowest method and about 15x that of the Illumina method. The most popular method as determined in section 4.2 (SWAN) is somewhere in the middle.

Preprocessing is something that only has to be performed once per cohort and the time taken for the procedure scales linearly with the number of probes in a chromosome, and individuals in the cohort. As such, computational time is less important for normalisation in comparison with other processing activities (such as calculating correlation matrices, which scales quadratically with the number of probes). This study has profiled the time taken to perform the normalisation procedure on the entire chromosome for a cohort of 120 individuals. Unless we are dealing with cohorts with several orders of magnitude more participants, or significantly more cohorts overall, the linear computational complexity and relative similarity of runtime (within one order of magnitude) between methods means that selection of normalisation type should not be particularly concerned with how long each method takes to run, and instead be primarily focused on results.

4.4 Study: Statistical differences in beta values due to selection of normalisation type

4.4.1 Rationale

Before investigating the differences between beta correlation matrices caused by normalisation type selection, we undertook an investigation of the effects it has on the beta values themselves. This study observes the effects of normalisation type on the statistical properties of beta values.

4.4.2 Methods

Beta matrices for each chromosome and normalisation method, as well as one containing CpG sites on all autosomal chromosomes, were computed for each normalisation method using the relevant steps of section 2.2. For all matrices, the mean and standard deviation for beta values at each CpG site were calculated using the *pandas* and *numpy* Python modules. The average mean and average standard deviation was then calculated for each matrix. A pair of 'offset' parameters (explained in section 2.2.4) were calculated as a difference between the calculated averages and:

- Autosome-wide averages for each normalisation method, so that relative differences between chromosomes can be assessed
- Chromosomal averages for each normalisation method, so that relative differences between normalisation methods can be assessed

The procedure was run using System 1 running Python configuration 1 as defined in Appendix A.

4.4.3 Results

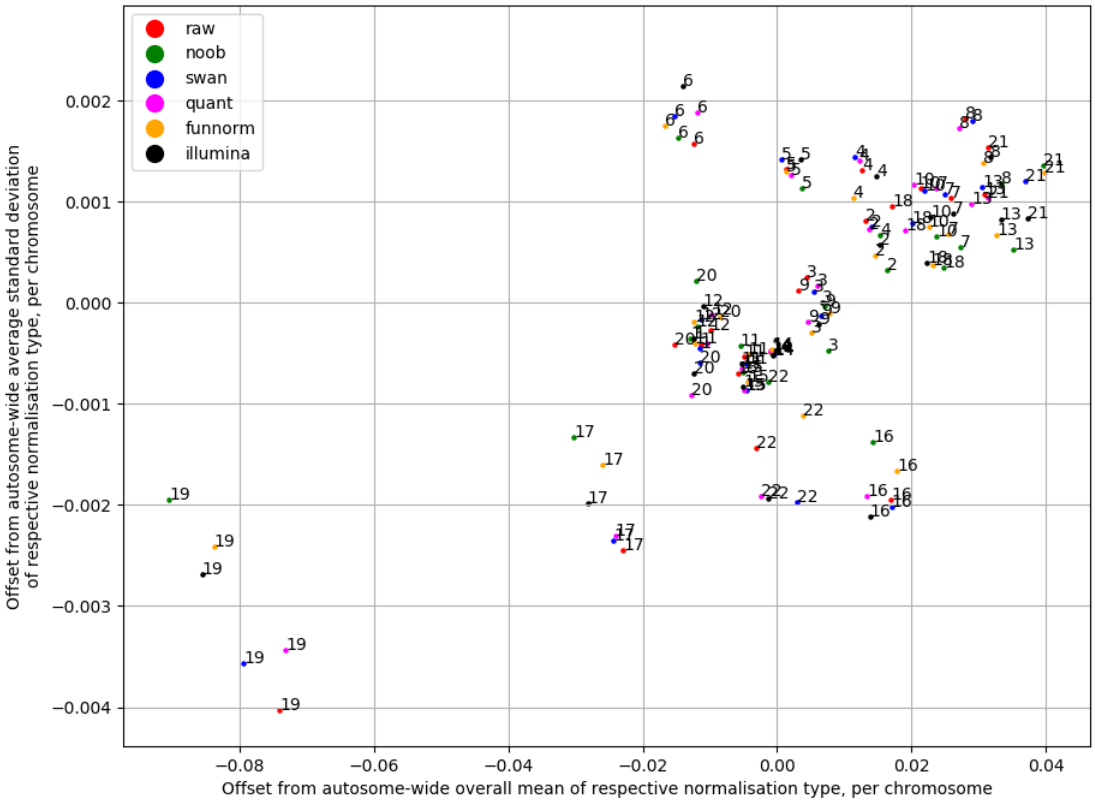

Figure 4.1: Differences between chromosome-average mean and all-autosome-average mean, and chromosome-average standard deviation and all-autosome-average standard deviation, for CpG methylation intensity (beta) values in each autosome. Source data: CHDS dataset

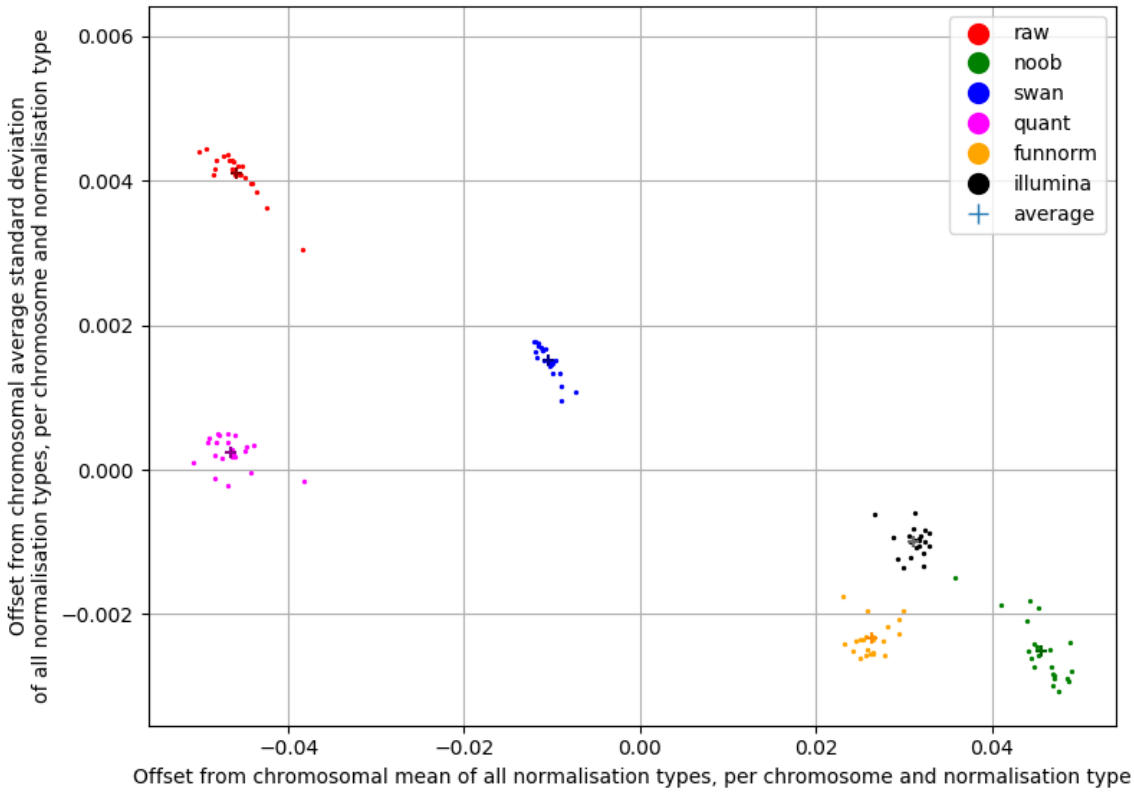

Figure 4.2: Differences between chromosome-average mean for a given normalisation type and chromosome-average mean of all normalisation types, and chromosome-average standard deviation for the given normalisation type and chromosome-average standard deviation of all normalisation types, for CpG methylation intensity (beta) values in each autosome. The average for each normalisation type is also included (marked with +); this average is unweighted so smaller chromosomes have a proportionally-greater effect. Source data: CHDS dataset

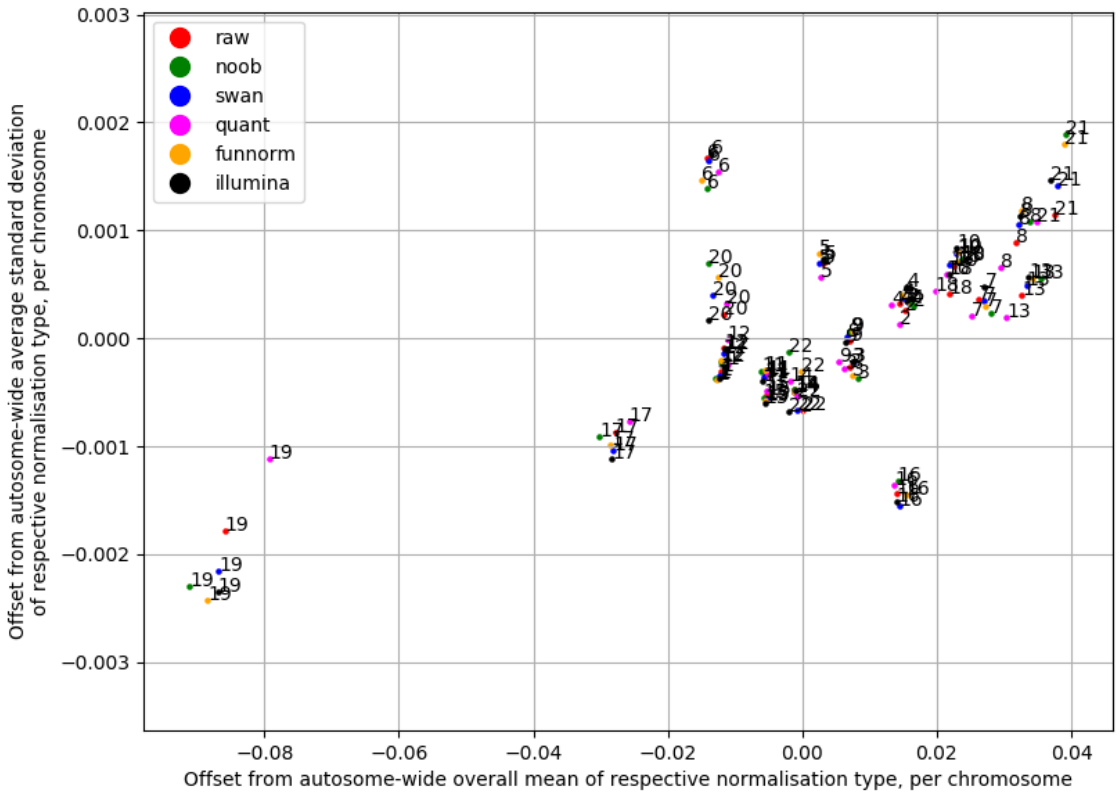

Figure 4.3: Differences between chromosome-average mean and all-autosome-average mean, and chromosome-average standard deviation and all-autosome-average standard deviation, for CpG methylation intensity (beta) values in each autosome. Source data: MTAB-7069 dataset

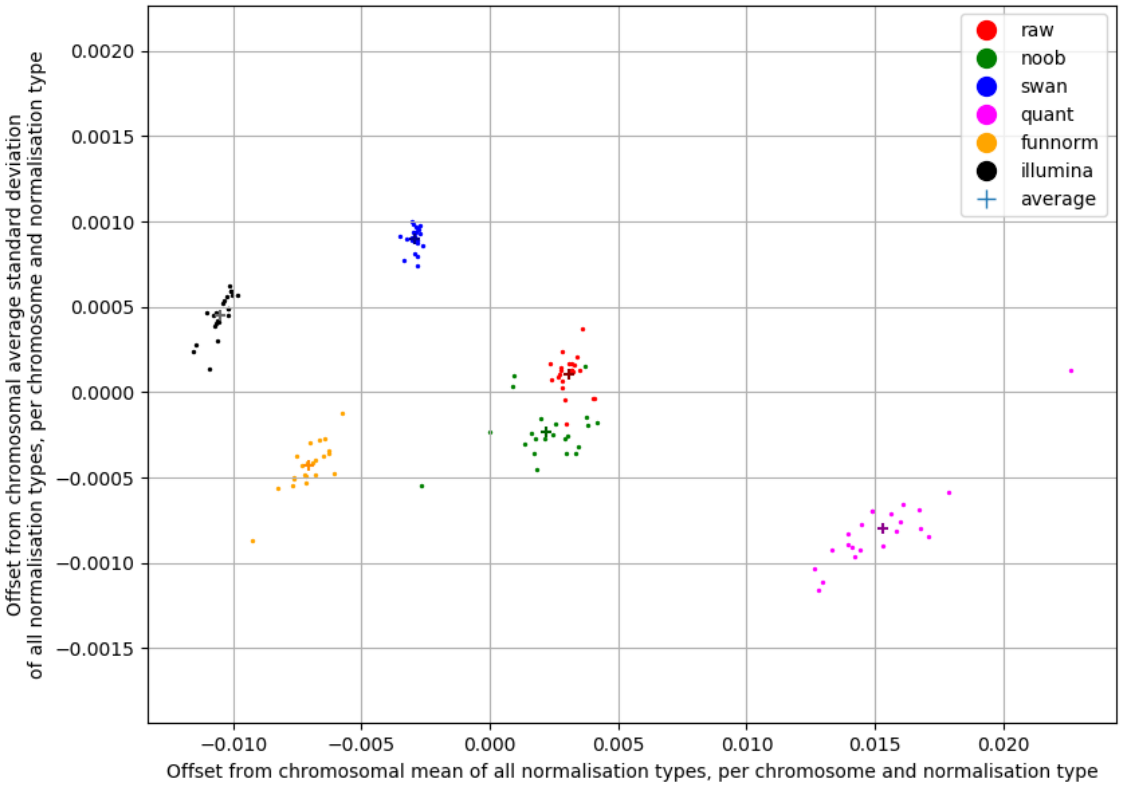

Figure 4.4: Differences between chromosome-average mean for a given normalisation type and chromosome-average mean of all normalisation types, and chromosome-average standard deviation for the given normalisation type and chromosome-average standard deviation of all normalisation types, for CpG methylation intensity (beta) values in each autosome. The average for each normalisation type is also included (marked with +); this average is unweighted so smaller chromosomes have a proportionally-greater effect. Source data: MTAB-7069 dataset

We can see that both cohorts present notable chromosome-based clustering for the plots of average standard deviation offset versus overall mean offset per normalisation type (figures 4.1 and 4.3). This particularly evident from chromosomes 6, 16, 17 and 19. General visual trends, such as a 'arrow'-shape are present (more visible in figure 4.3), and relative positions of the centres of these clusters are positioned similarly.

The same offsets per chromosome also show a distinct pattern of clustering (4.2 and 4.4), though the cluster placement does not appear to be as consistent.

4.4.3.1 Allosomes for comparison

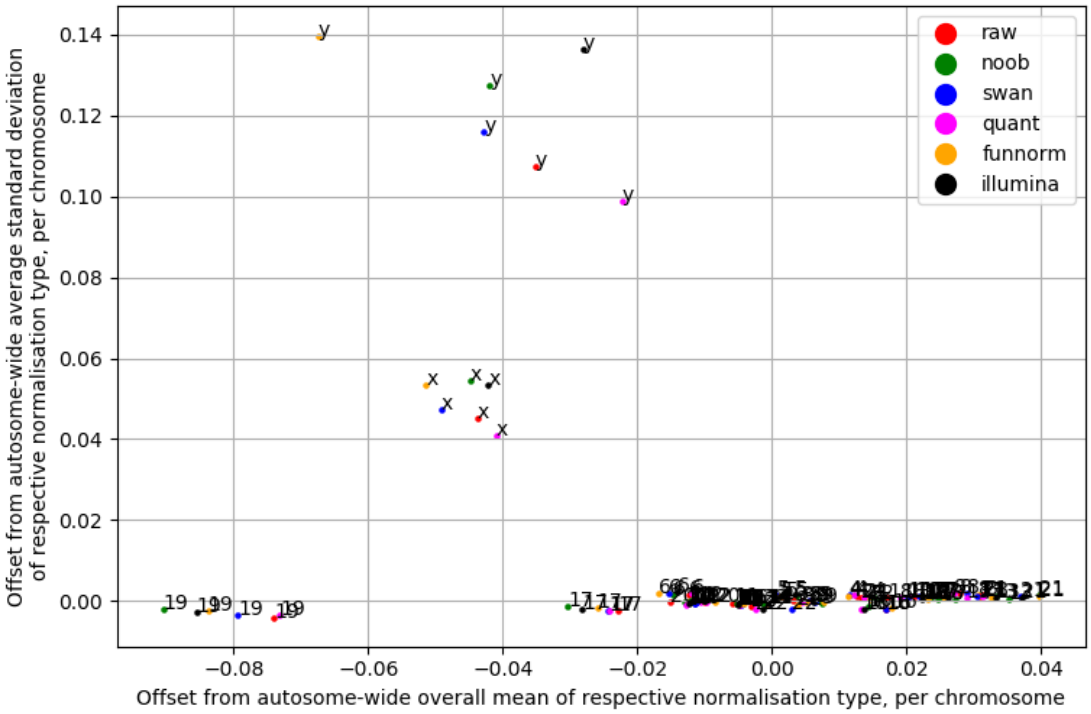

Figure 4.5: Differences between chromosome-average mean and all-autosome-average mean, and chromosome-average standard deviation and all-autosome-average standard deviation, for CpG methylation intensity (beta) values in each chromosome, including sex chromosomes (which are not included in the all-autosome averages). Source data: CHDS dataset

The sex chromosomes were excluded from the analysis in section 4.4.3 as most individuals will either have a Y chromosome or experience X-inactivation, and this complicates epigenetic studies. We plot the offsets of the X and Y chromosome for the CHDS dataset in figure 4.5. We can see that chromosomes X and Y have significantly higher average standard deviation for all normalisation methods.

4.5 Study: Statistical differences in beta correlation matrices due to selection of normalisation type

4.5.1 Rationale

Beta values themselves can vary significantly in statistical properties due to selection of array normalisation method, as shown in section 4.4 and discussed in section 4.6. To investigate whether underlying epigenetic trends may still be extractable from the data, we can compare their correlation matrices. It is hypothesised

that those trends will still be present (at least partially) and that this will manifest as an overlap in 'strong' correlations between beta correlation matrices produced with different normalisation methods.

4.5.2 Methods

Beta correlation matrices for all six normalisation types (including raw) were calculated for the CHDS dataset using methods described in section 2.2. Statistical analysis of these correlation matrices was conducted using the *pandas* and *numpy* Python libraries where possible. It was only possible to analyse correlation matrices up to a certain size, owing to technical limitations of the HPC resource used for the study (discussed in section 2.3.1). Nonetheless, we were able to use available HPC resources to analyse a subset of chromosome beta correlation matrices and compare their mean and standard deviation. This procedure was repeated with the MTAB-7069 dataset (described in appendix B) to confirm that normalisation methods have similar effects on beta correlation matrices for other cohorts (the caveats regarding use of this cohort are described in section 4.6). Chromosomes available for this study include: 4, 8, 9, 13, 14, 15, 16, 18, 20, 21, 22, X and Y. This subset of chromosomes provides coverage of half of the autosomes in the human genome, as well as both allosomes. It should be noted that this subset does not provide coverage of half of all probed CpG sites, but as we are focused on chromosomes specifically, this is accepted as a limitation for this study.

Similar to methods described in section 4.4.2, we calculate an overall mean and variance for each correlation matrix (of which there is one for every combination of chromosome and normalisation type). To compare normalisation types, offset parameters for mean and variance are calculated as the difference between the actual values for a correlation matrix and the average of values across all normalisation types for a given chromosome. From this, we can identify the relative 'shift' in mean and variance occurring as a result of normalisation method choice.

Additionally, the beta correlation matrices (abbreviated as BCMs) for several chromosomes are analysed in depth. Chromosomes for these matrices include 21, 22, X and Y. The following metrics were ascertained for each matrix:

- Overall average, variance, skew and excess kurtosis
- Average positive correlation
- Average negative correlation
- Ratio of positive to negative correlations

We also determined the strongest 10% positive and negative correlations for each beta correlation matrix. From this, we identified which pairs are considered strong (per chromosome) across all six methods, and which pairs are considered strong in only one method. This information is used later to compare the different methods of normalisation.

4.5.3 Results

4.5.3.1 Beta correlation matrix offsets in mean and variance

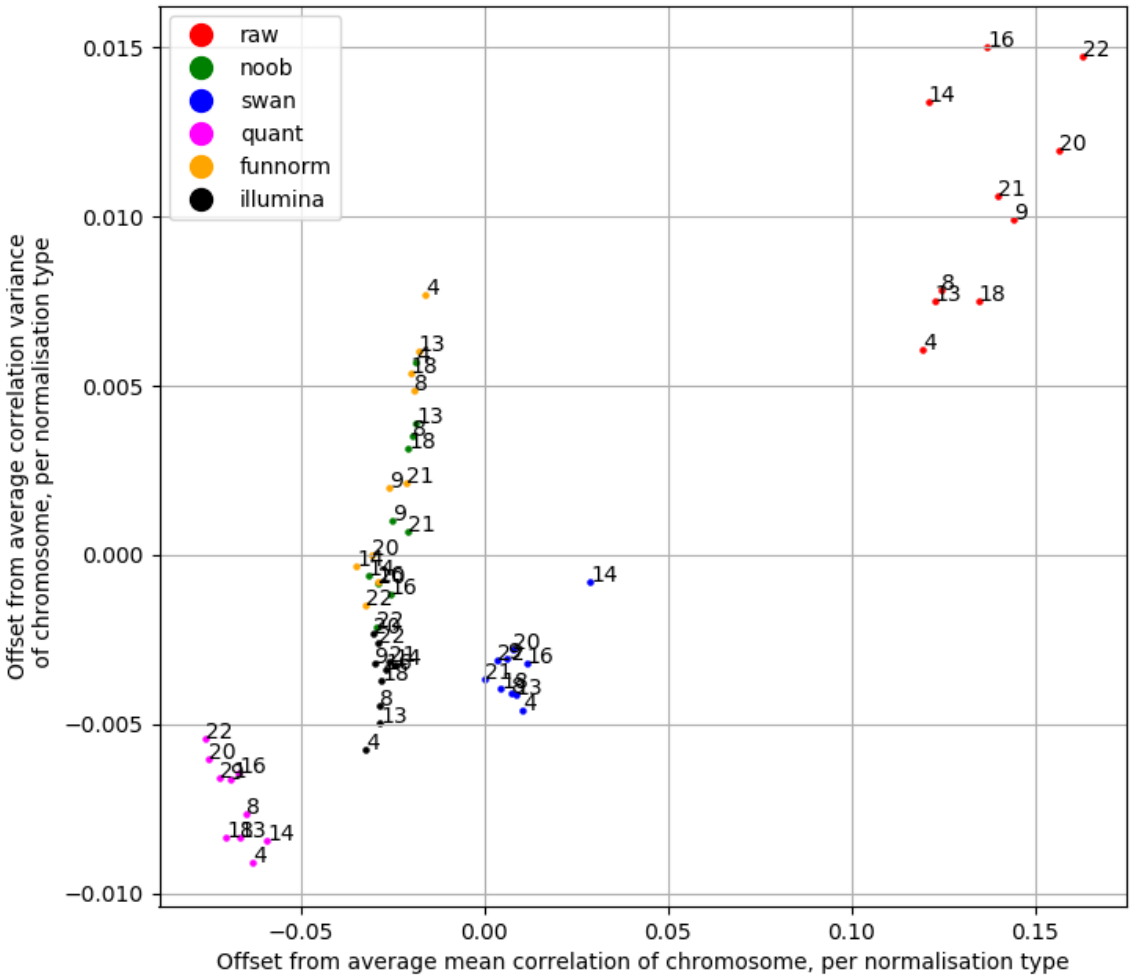

Figure 4.6: Differences between beta correlation matrix (BCM) mean and average of BCM means for all norm types, and BCM variance and average of BCM variances for all norm types, for selected chromosomes. Source data: CHDS dataset

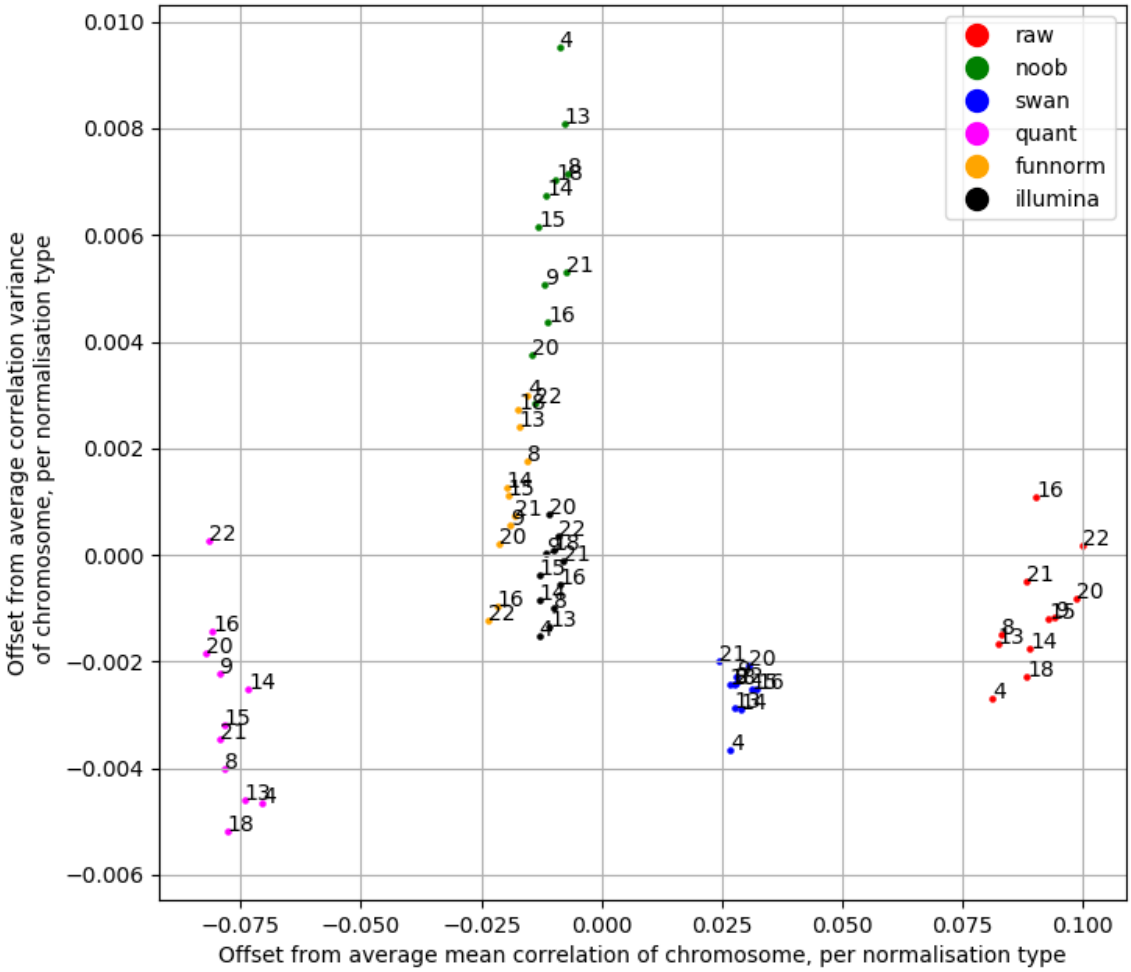

Figure 4.7: Differences between beta correlation matrix (BCM) mean and average of BCM means for all norm types, and BCM variance and average of BCM variances for all norm types, for selected chromosomes. Source data: MTAB7069 dataset

Strikingly, both separate datasets produce a similar clustering based on normalisation-type, as shown in figures 4.6 and 4.7. In both cases, quantile normalisation tends to have the lowest offset in average mean, followed by NOOB, functional normalisation and Illumina normalisation (in the same cluster), then SWAN, then the unnormalised set. In terms of variance, there is significantly more overlap in the range of values for each normalisation type.

4.5.3.2 Statistical analyses of beta correlation matrices for chromosome 21

	raw	noob	swan	quant	funnorm	illumina
Overall average	0.214	0.0531	0.0739	0.00194	0.0526	0.0482
Overall variance	0.0336	0.0237	0.0193	0.0164	0.0252	0.0199
Overall skew	-0.103	0.344	0.157	0.158	0.427	0.121
Overall excess kurtosis	-0.195	2.06	1.73	2.66	1.88	1.89
Average positive correlation	0.249	0.134	0.136	0.0986	0.137	0.124
Average negative correlation	-0.0833	-0.0919	-0.0789	-0.0951	-0.0937	-0.0897
Ratio of positive to negative correlations	6.69	1.77	2.48	1.01	1.71	1.79

Table 4.1: Statistical comparison of beta correlation matrices of Chromosome 21 for different normalisation methods (CHDS cohort)

Overall averages and variances are highest for the BCM produced from unnormalised data, and lowest for that produced from data normalised using Quantile normalisation. A negative skew for the raw data BCM suggests the centre of mass is shifted to the right, while positive skews for all other normalisation types suggests the opposite. Similarly, the raw data BCM is platykurtic (negative excess kurtosis), implying a broader distribution with thinner tails; the other normalisation types are all leptokurtic (positive excess kurtosis) which suggests fatter tails and a denser centre. In all cases, the average positive correlation has a greater magnitude than the average negative correlation. There also tend to be more positive correlations than negative correlations; all normalisation types have a positive ratio of positive to negative correlations, though the quantile-normalised BCM is very close to one. The quantile-normalised BCM also has the lowest average variance within each CpG site.

4.5.3.3 Statistical analyses of beta correlation matrices for chromosome 22

	raw	noob	swan	quant	funnorm	illumina
Overall average	0.243	0.0508	0.0839	0.00438	0.0481	0.0513
Overall variance	0.0381	0.0212	0.0202	0.0179	0.0218	0.0208
Overall skew	-0.172	0.267	0.202	0.165	0.268	0.205
Overall excess kurtosis	-0.407	2.24	1.57	2.3	1.87	1.67
Average positive correlation	0.277	0.127	0.143	0.104	0.128	0.128
Average negative correlation	-0.0836	-0.0891	-0.0778	-0.0982	-0.0916	-0.0898
Ratio of positive to negative correlations	7.37	1.81	2.75	1.04	1.73	1.8

Table 4.2: Statistical comparison of beta correlation matrices of Chromosome 22 for different normalisation methods (CHDS cohort)

The relative statistical properties of beta correlation matrices for chromosome 22 are very similar so the commentary in section 4.5.3.2 is applicable here as well. In short:

- Overall average and variance is highest for the BCM of the unnormalised data

- Skew and kurtosis are both negative for the unnormalised BCM, and positive for all other normalisation types
- The magnitude of average positive correlation is greater than that of the average negative correlation in all cases
- The ratio of positive to negative correlations is greater than one in all cases, but only marginally-so for Quantile-normalised data

4.5.3.4 Statistical analyses of beta correlation matrices for chromosome X

	raw	noob	swan	quant	funnorm	illumina
Overall average	0.0972	0.0362	0.0452	0.000295	0.0476	0.0155
Overall variance	0.152	0.176	0.157	0.157	0.177	0.167
Overall skew	-0.0326	-0.0762	0.0226	0.000889	-0.104	-0.0257
Overall excess kurtosis	-1.22	-1.41	-1.23	-1.39	-1.41	-1.36
Average positive correlation	0.395	0.391	0.369	0.35	0.394	0.367
Average negative correlation	-0.276	-0.355	-0.312	-0.35	-0.354	-0.351
Ratio of positive to negative correlations	1.4	1.15	1.16	1.0	1.19	1.07

Table 4.3: Statistical comparison of beta correlation matrices of Chromosome X for different normalisation methods (CHDS cohort)

The beta correlation matrices for chromosome X have different statistical tendencies than those for the autosomal chromosomes we tested. While the trend of the raw BCMs having a higher overall average is maintained, we see significant differences in most other metrics. Overall variance does not differ much between normalisation types, and we can see that in all cases, the variance is in the next order of magnitude compared with the autosomal chromosomes. Skew is no longer negative for only the raw BCMs, and excess kurtosis is negative for all normalisation types. As in the autosomes, the ratio of positive to negative correlations tends to be above 1.0 (except for Quantile normalisation which is very close to one) and the magnitude of average positive correlations tends to be higher than that of the average negative correlation (again, with the exception of Quantile normalisation).

4.5.3.5 Statistical analyses of beta correlation matrices for chromosome Y

	raw	noob	swan	quant	funnorm	illumina
Overall average	0.14	0.103	0.0993	0.0564	0.102	0.0604
Overall variance	0.195	0.231	0.194	0.196	0.214	0.12
Overall skew	-0.261	-0.226	-0.224	-0.189	-0.354	-0.172
Overall excess kurtosis	-1.46	-1.54	-1.52	-1.59	-1.54	-1.19
Average positive correlation	0.462	0.464	0.439	0.414	0.443	0.316
Average negative correlation	-0.337	-0.4	-0.363	-0.397	-0.418	-0.286
Ratio of positive to negative correlations	1.54	1.34	1.38	1.25	1.47	1.35

Table 4.4: Statistical comparison of beta correlation matrices of Chromosome Y for different normalisation methods (CHDS cohort)

As with chromosome X, BCMs for chromosome Y generally show different statistical than those for the autosomes that we tested. Overall average is once again higher for the unnormalised BCMs than for any of those using normalised data, but variance does not show the same behaviour. As with chromosome X, our measure of kurtosis is consistently negative, though only in the case of chromosome Y is skew consistently negative for all normalisation methods. Like what we've mostly seen for the other chromosomes, the average positive correlation has a higher magnitude than the average negative correlation, and the ratios of positive to negative correlations are higher, for all normalisation methods.

4.5.4 Overlapping strong correlations

Chromosome	21	22	X	Y
Observed positive correlations	5,304,006	16,866,512	18,220,536	14,393
Consistently-correlating pairs	1,222,948	4,233,544	6,821,248	2,190
Unique positive correlations - raw	1,821,513	5,181,375	1,608,290	2,457
Unique positive correlations - NOOB	212,119	786,984	1,053,037	3,054
Unique positive correlations - SWAN	967,257	3,803,112	1,246,056	681
Unique positive correlations - quant	1,518,447	4,426,411	2,194,822	2,028
Unique positive correlations - FunNorm	435,388	848,841	2,246,787	2,741
Unique positive correlations - Illumina	531,024	1,086,739	783,833	3,142

Table 4.5: Overlapping positive correlations between selected normalisation methods, for chromosomes 21, 22, X and y (CHDS cohort)

Chromosome	21	22	X	Y
Observed negative correlations	5,304,006	16,866,512	18,220,536	14,393
Consistently-correlating pairs	814,694	2,274,764	6,098,384	8,975
Unique negative correlations - raw	1,712,001	6,023,526	1,840,373	816
Unique negative correlations - NOOB	262,143	825,236	591,885	930
Unique negative correlations - SWAN	659,721	2,434,576	644,893	355
Unique negative correlations - quant	1,794,978	5,932,545	2,722,840	2,524
Unique negative correlations - FunNorm	397,326	1,216,331	1,071,097	1,963
Unique negative correlations - Illumina	500,854	1,499,322	481,850	2,387

Table 4.6: Overlapping negative correlations between selected normalisation methods, for chromosomes 21, 22, X and y (CHDS cohort)

Tables 4.5 and 4.6 both show that there is a low degree ($\sim 10-20\%$) of overlap in strong correlations between all six methods. The raw and quantile methods produce the highest number of unique strong correlations for the autosomes, within 50% of an order of magnitude of each other. The quantile method produces the highest number of unique strong correlations for the allosomes, though functional normalisation also produces a lot of unique strong correlations. Conversely, we see that NOOB produces the fewest unique strong correlations for both autosomes. For allosomes, SWAN produces the fewest unique strong correlations in both cases.

4.6 Discussion

4.6.1 The effect of normalisation type on beta values for autosomes

Figure 4.1 shows some prominent chromosome-based clusters in this dataset. For example, chromosome 19 has a notably lower average mean and standard deviation for its associated beta values, which form their own 'island' on the graph. Other notable islands exist for chromosomes 5, 6, 16 and 17, and results for the six normalisation methods tend to be relatively co-located for most of the chromosomes. This suggests that the choice of normalisation method doesn't drastically alter the statistical distribution of beta values, relative to other methods; the difference is driven largely by the chromosomes themselves.

Figure 4.2 shows statistical difference based on normalisation type. An intuitive way of thinking about it would be considering one of the clusters in figure 4.1, and taking the difference between each point and the mean of the cluster. The clusters in figure 4.2 are very well-distinguished and show a clear demarcation between the different normalisation types. Of particular note is that the case of no normalisation ('raw' in the figure) tends to produce data with a higher standard deviation than any of the other methods. This is possibly due to other methods removing some of the 'noise' in the data, whether intentionally or not.

For comparison, the same procedure was run for the MTAB-7069 dataset (also in appendix B). The results can be seen in figures 4.3 and 4.4. It must be noted that the MTAB-7069 dataset is rarely-used in this thesis owing to the small size of its cohort and the presence of confounding factors due to taking multiple samples from the same individuals. The MTAB-7069 has 11 participants and takes three blood samples from each (one from the umbilical cord at birth, one at 5 years and one at 10 years) for a total of 33 samples. Influences from both genetic factors and age, discussed in section 1.1, are likely to be significant confounding

factors in this dataset. Nonetheless, we felt it would be appropriate to test the performance of our six normalisation methods on this dataset as well.

Figures 4.1 and 4.3 show a very similar trend in offsets per chromosome. Many of the clusters present in the MTAB-7096 chromosome offsets have a similar position relative to each other as the clusters in the CHDS chromosome offsets. This suggests that the overall mean DNA methylation and average standard deviation of this at each CpG site tends to be chromosome-specific.

Figures 4.2 and 4.4 do not show any concordant relationship. Clusters are present in both but their relative positions do not appear to follow a specific trend. A potential explanation for this is inconsistent selection of reference probes by normalisation algorithms (for methods that use reference probes). Unfortunately, the previously-observed high relative standard deviation offset of the non-normalised betas is not reproduced in this dataset.

This study suggests that the statistical properties of beta values can vary greatly depending on which normalisation type is selected, and the effects of a particular normalisation type are not necessarily consistent between cohorts. This poses a problem for meta-analyses or new research that combines beta values from past studies, and use these beta values directly - additional care must be taken to ensure normalisation type is consistent between cohorts when combining pre-processed data, or researchers must obtain unprocessed data from which they can generate beta values with consistent normalisation. Studies in this thesis do not combine cohorts, so we do not face this problem; however, we discuss some theory regarding cohort combination in section 8.5.9.

4.6.2 The effect of normalisation type on beta values for allosomes

CpG sites on the X and Y chromosomes have a significantly higher average standard deviation than any of the autosomal chromosomes, as per their very high offsets of average standard deviation shown in figure 4.5. It is proposed that this is due to X-inactivation in the case of the X chromosome (described in section 1.2.3), as this process is associated with gains in methylation at a number of silenced genes (Sharp et al. 2011) so sites that would typically experience lower levels of methylation on the active X chromosome would be significantly more-methylated on the inactive chromosome. For the Y chromosome, it is suggested that this may be due to calculation of beta values of this chromosome for individuals who do not have one, as the cohort consists of males and females (by gender) and the beta matrix for the Y chromosome does not exclude samples from anyone. This is perhaps something that should be considered for later analyses; given the difficulty of interpreting data derived from allosomes, then maybe they should be considered separately or omitted from analyses where appropriate. The major factors that play into X- and Y- chromosome regulation are biological sex, and if we intend to obtain results applicable to people regardless of this, then we will need to devise methodologies specific to these chromosomes. We approach later studies in this thesis with this fact in mind, opting to separate out allosomal data from autosomal data as required.

4.6.3 The effect of normalisation type on beta correlation matrices

4.6.3.1 Statistical properties

Figures 4.6 and 4.7 show the effects of normalisation type on the mean and variance of the beta correlation matrix, for CHDS and MTAB7069 cohorts respectively. Both figures show four primary clusters, positioned similarly (relative to other clusters) for both cohorts. From left to right (or smaller to larger mean):

- The Quant (stratified quantile normalisation) cluster
- The NOOB-FunNorm-Illumina cluster
- The SWAN cluster
- The raw cluster

The raw cluster has the greatest positive offset in mean; tables 4.1 to 4.4 also show this as the overall average for the raw method is more positive that of all other methods for the four chromosomes analysed in detail. It is proposed that this is due to the background fluorescence signal which remains unaccounted for in this method. This increase in average correlation may make it harder to delineate strong correlations with threshold-selection and may introduce noise which makes it more difficult to find them with proportional selection.

The Quant cluster has the the greatest negative offset in mean. Tables 4.1 to 4.4 show that the overall average for the Quant method is closer to zero than those of other methods for the four chromosomes analysed in detail. We also see that the ratio of positive to negative correlations is the closest to 1.0 in all of these chromosomes. This suggests that the method produces a very balanced correlation matrix. Beta correlation matrix variance was lowest for the Quant method for the autosomes, though this wasn't the case for the allosomes. We also see that the magnitude of excess kurtosis for the Quant beta correlation values tends to be high for autosomal chromosomes, relative to the other normalisation types. This suggests that the distribution is 'squished' inwards, making the tails of the distribution more prominent. This would make it easier to select a threshold for strong correlations, as there is a broader 'surface' along which we can set the threshold without dramatically increasing the number of correlations we need to consider. We could therefore use the high relative kurtosis in beta correlation matrices produced by the Quant method to our advantage. This is something we should keep in mind if we have issues with threshold selection.

The NOOB-FunNorm-Illumina cluster, in conjunction with tables 4.1 to 4.4, shows that these three methods had a similar overall mean. Indeed, the aforementioned tables also show that they are similar in other statistical properties, particularly for the autosomes - in particular, their ratio of positive to negative correlations was very similar in both cases (within 0.1) whereas the other methods produced ratios with a greater difference than this. It is assumed that these three methods produced similar beta correlation matrix means due to having a similar effect on the beta values themselves - figure 4.1 shows that these three methods had a comparable mean offset clustering for the beta values prior to correlation.

It is difficult to suggest a normalisation method on the basis of the statistical properties of its correlation matrix. In the absence of past studies suggesting the contrary, we assume that the NOOB, FunNorm and Illumina methods are the best options when evaluating beta correlation matrices, as their most basic

statistical properties are the most similar - we can take this as suggesting that other methods have ramifications which alter correlation matrices in a negative way. This is not necessarily the case, as it is just as likely that these three methods are equally bad, rather than good, so further information must be considered when selecting a normalisation method to use in future studies. To that end, we will need to make a decision based on other information - in this case, overlapping strong correlations. The outcome of our discussion on strong correlation overlaps (section 4.6.3.2) applies to correlation studies in general as we will typically be interested in these strong correlations more than anything else.

4.6.3.2 Strong correlation overlaps

Tables 4.5 and 4.6 both show that there is a low degree ($\sim 10 - 20\%$) of overlap between strong correlations of all six methods. Though it may be small, this overlap could be indicative of genuinely strong and biologically-meaningful relationships, as they're present regardless of what we do to the data prior to correlating. In future studies, we could consider using multiple methods when assessing the strength of correlations to improve the robustness of results. An approach to this, which we refer to as *metaN*, is described later in this discussion (section 4.6.4.1).

We see that the raw and Quant methods produce the highest number of unique strong correlations for the autosomes. A high number of unique correlations could suggest one of two things:

1. the method is well-suited to finding biologically-meaningful correlations, and all other methods are not
2. the method is finding correlations that are occuring only due to the method, and this is not biologically-meaningful

If the first of these is correct, then we would see one method with a high number of unique strong correlations, and all of the others would have a low number. This is not the case, and as the second of the above can apply to multiple methods, it is more likely to be what is happening in this instance; by this logic, the raw and Quant methods are ill-suited to finding biologically-meaningful correlations for these chromosomes in the CHDS dataset. We can expand upon the above logic to identify the best-performing method for strong correlation identification under the conditions in this study. In the absence of any evidence to the contrary, we have to initially assume that all methods are generally equivalent. If there is not a single 'stand-out' method with a high number of unique strong correlations, then we should instead consider the method with the fewest unique strong correlations. This is because this method will produce the highest number of correlations that are backed up by at least one other method.

For autosomes, we see that NOOB produces the fewest unique strong correlations in both cases. For allosomes, SWAN produces the fewest unique strong correlations in both cases. Based on this, we should consider using both NOOB and SWAN in correlation studies where identification of strong correlations using a single normalisation method is important, with NOOB being used preferentially for autosomes and SWAN being used for allosomes.

4.6.4 Development of combination methods

Each of the methods described in section 4.1 have their advantages and disadvantages, and all have been used individually in studies by other research. In the context of finding correlating CpG sites - if we're not

too concerned about the relative strength of correlations, just whether or not we consider them to be above some arbitrary threshold, then we can take the approach of using multiple sets of betas as calculated with different normalisation methods (from the same source data). For example, if several different normalisation methods all suggest that a specific pair of CpG sites tend to correlate strongly, then the likelihood that there is an underlying association is higher. To that end, we can consider 'meta' methods to make use of results from multiple different normalisation methods.

4.6.4.1 metaN

The most straightforward meta method would be to require that a correlation is present in some number of normalised datasets. We refer to this as *metaN*, where N is the minimum number of methods that yield a strong correlation for the meta method to consider a correlation to be genuinely strong. For example, we can define *meta5* as the following:

- meta5: a strong correlation only exists between the methylation intensities of two CpG sites if all five of the selected methods of normalisation (NOOB, Quant, FunNorm, SWAN, Illumina) produce a strong correlation between the two sites

We can also define *meta6* as having the same criteria as *meta5*, except with the additional criterion that a strong correlation must also exist in the unnormalised data. *meta1* is the trivial case where we consider a correlation to be genuinely strong if it is present following any method of normalisation (including not normalising).

We can also define a 'metaN score' for a given pair of CpG sites, which is the number of methods that yield a strong correlation between these sites. In this case, a score of six suggests that it was found in all methods, and conversely, a score of zero suggests that no method identified a strong correlation between these CpG sites. A higher score increases the likelihood that a biologically-meaningful association is present.

These methods do not provide a derived correlation coefficient, but rather are a means of suggesting a more significant association. If a coefficient is required, then it can be (for example) taken as an average of each of the individual N correlations, or the lowest correlation can be taken, etc. This will be defined as required on a per-study basis.

4.7 Concluding remarks

In this chapter, we have investigated the effects of normalisation method choice on the resulting Spearman beta correlation matrices, derived from several chromosomes, using the CHDS and MTAB7069 cohorts. Our findings suggest that there are certainly differences in the outputs of each method, and this needs to be taken into account when deciding which normalisation method to apply to a dataset.

Both NOOB normalisation and no normalisation (raw) were used in the preliminary study of chapter 3, for reasons described in section 3.2. The results of studies in this chapter have supported continued use of NOOB normalisation prior to calculation of correlation matrices, but only for autosomes. SWAN should be considered for allosomes owing to its better performance in this study. In both cases, their use is recommended on the basis that their strong correlations are more likely to be backed up by at least one other

method; in the absence of a single method that picks up the majority of unique strong correlations, we instead assume that all methods have some merit, and that the method with the fewest unique strong correlations is best at capturing biologically-meaningful correlations. In the instance where a single method is insufficient, we can instead use the metaN method to identify strong correlations using a scoring system, where the score of a correlating pair is equal to the number of methods in which it is identified a as strong, given a consistent threshold or proportion for strength identification.

Further studies could validate results in this section using more chromosomes, different test cohorts, etc. For the time being, we consider there to be sufficient evidence to continue to use the NOOB and SWAN normalisation methods when evaluating correlations in methylation intensity where use of a single method is required. When we are not constrained to a single method, use of the *metaN* approach should be considered as it may provide better results.

Chapter 5

A comparison of different correlation methods

5.1 Premise

A correlation is a statistical relationship between two variables. The overarching theme of research in this thesis involves identifying CpG sites on the genome that correlate in methylation intensity. It is therefore important that we evaluate our different options for calculating correlation coefficients. As discussed in section 1.3.2, we have three common measures:

- Pearson correlation
- Spearman rank correlation
- Kendall rank correlation

The preliminary study in chapter 3 used the Spearman method and we were able to generate a basic correlation network from the results. However, we should also quantify the performance of all three of our possible measures of correlation, so we can make better-informed decisions for future studies.

In this chapter, we compare the different methods of calculating correlation and some considerations regarding their use. Two studies were undertaken as part of an effort to assess correlation method. The first was computational profiling to characterise the performance of the three correlation coefficients, and the second was a statistical analysis comparing the results.

5.2 Study: Computational considerations

To compare the time taken to generate beta correlation matrices, the three methods were subjected to computational profiling. Correlation matrices consist of an n-by-n matrix for a set of n variables. The total number of unique values that need to be calculated for these matrices is $\frac{n(n-1)}{2}$, so the number of correlation coefficients that have to be calculated grows quadratically with the size of the dataset. This poses a problem for DNA methylation data, as the more CpG sites we look at (and thus more values of methylation intensity from which we have to calculate correlations), the more computationally-intensive the procedure becomes.

To ensure that correlation matrices can be calculated in an acceptable time frame, we need to investigate how long they take to generate on an appropriate dataset.

5.2.1 Methods

Raw data was processed and correlations generated as per the methods discussed in section 2.2, for the CHDS cohort. The serialisation step was included as it is a key component of the pipeline and should take the same period of time regardless of method. Chromosome 21 was selected for profiling as it is the smallest autosomal chromosome, with 10300 probes on the EPIC array.

Profiling was undertaken by measuring the runtime of a Python function (using the standard *datetime* library) containing the following processing steps:

1. Reading the CSV file containing beta values for a specific normalisation method and chromosome
2. Calculating the beta correlation matrix for a specific correlation type
3. Serialising the correlation matrix

This was repeated separately for each correlation type so profiling measured the same high-level operations for each method. Profiling was undertaken using System 2 and Python configuration 2 as described in Appendix A. Correlations are generated via the implementation in the Python *pandas* library as this is the main data processing library we use in this thesis.

5.2.2 Results

	Spearman	Pearson	Kendall
Run 1 (s)	195.04	27.87	20652.82
Run 2 (s)	202.24	35.97	19923.58
Run 3 (s)	203.63	34.77	19907.44
Average	200.30	32.87	20161.28

Table 5.1: Computational profiling of selected correlation methods: time taken (in seconds) to generate beta correlation matrices for Chromosome 21.

All three of our trials show the same results - the Spearman method is roughly an order of magnitude slower than the Pearson method, and the Kendall method is roughly two orders of magnitude slower than the Spearman method. These results include all steps as described in section 5.2.1.

5.3 Study: Comparison of coefficients calculated between methylation intensities of CpG sites, for selected chromosomes

To compare results for each of the three correlation methods, they were also subjected to a statistical analysis. While no literature exists on the topic of correlations in DNA methylation intensity (and consequently, we have no benchmark to compare against) we may be able to gain some understanding of the benefits and downsides of each method by statistically comparing the resulting correlation matrices. As the precursor

data for each chromosome (in terms of the beta values) used for each different correlation method is the same, any variation in statistical quantities such as mean, variance etc. is entirely due to the correlation method, as opposed to the data. As such, intensive statistical tests are not required and we can simply use a direct comparison. Given the comparatively-extreme length of time required to generate Kendall correlation matrices, only a small selection of chromosomes will be used in this analysis.

5.3.1 Methods

Raw data was processed with NOOB preprocessing and correlations generated as per the methods discussed in section 2.2, for the CHDS cohort. Chromosomes 21, X and Y were selected for comparison, on the following bases:

- Chromosome 21 is the smallest autosomal chromosome (10300 probes)
- Chromosomes X (19090 probes) and Y (537 probes) were shown to have a significantly higher standard deviation than autosomal chromosomes in section 4.4.3.1, and the possibility that this increase in standard deviation may be dependent on correlation method is something that should be investigated.

The *pandas* and *scipy* modules were used for statistical analysis.

For overlap comparison, a proportional threshold (described in section 2.2.5.2) was used - the strongest 10% positive and negative correlations were taken from each correlation matrix. We compare the number of strong positive and negative correlations that are found throughout all three methods ('consistent correlations'), and the number found only by one method ('unique correlations'), with the total number of strong positive and negative correlations.

This study used system 2 and Python configuration 2 (as described in Appendix A).

5.3.2 Results

5.3.2.1 Statistical comparisons

Correlation type	**Spearman**	**Pearson**	**Kendall**
Mean correlation	0.05314	0.05319	0.03622
Mean variance of correlation	0.02183	0.02274	0.01031
Mean positive correlation	0.13389	0.13503	0.09093
Mean negative correlation	-0.09187	-0.09239	-0.0622
Ratio of positive to negative correlations	1.772	1.751	1.776

Table 5.2: Statistical comparison of different correlation methods, applied to the NOOB-normalised betas for Chromosome 21 (10300 sites)

Correlation type	Spearman	Pearson	Kendall
Mean correlation	0.03625	0.10465	0.04710
Mean variance of correlation	0.17609	0.42183	0.08872
Mean positive correlation	0.39069	0.58791	0.27996
Mean negative correlation	-0.35499	-0.53070	-0.25123
Ratio of positive to negative correlations	1.146	1.128	1.154

Table 5.3: Statistical comparison of different correlation methods, applied to the NOOB-normalised betas for Chromosome X (19090 sites)

Correlation type	Spearman	Pearson	Kendall
Mean correlation	0.10272	0.1364	0.07669
Mean variance of correlation	0.19755	0.51011	0.09372
Mean positive correlation	0.46421	0.71292	0.32278
Mean negative correlation	-0.40037	-0.64829	-0.26800
Ratio of positive to negative correlations	1.342	1.298	1.345

Table 5.4: Statistical comparison of different correlation methods, applied to the NOOB-normalised betas for Chromosome Y (537 sites)

The Spearman rank correlation matrix and the Pearson correlation matrix had similar properties for chromosome 21 (mean correlation, mean variance of correlation, mean positive correlation, mean negative correlation and ratio of positive to negative correlations all within 5%), compared with the Kendall matrix which was notably different for all of our metrics except ratio of positive to negative correlations.

The same degree of similarity was not found for either of the sex chromosomes; while the ratio of positive to negative correlations was within 10% in all cases, other statistical properties varied substantially compared to the patterns we saw in chromosome 21.

The only consistent trend we see across all chromosomes and correlation methods is that the ratio of positive to negative correlations is always greater than one, and the magnitude of the mean positive correlation is always greater than the magnitude of the mean negative correlation.

5.3.3 Overlaps and uniqueness of strong correlations

The displayed number of positive correlations for each chromosome does not include the trivial values of 1.0 (for self-correlation).

Chromosome	21	x	y
Observed positive correlations	5,303,985	18,220,451	14,392
Consistent positive correlations	4,385,976	8,570,169	6,081
Unique positive correlations - Pearson	808,822	8,935,654	7,931
Unique positive correlations - Spearman	81,425	326,712	503
Unique positive correlations - Kendall	81,771	621,339	424
Observed negative correlations	5,303,985	18,220,451	14,392
Consistent negative correlations	4,142,041	10,724,161	7,114
Unique negative correlations - Pearson	1,040,737	6,718,785	6,716
Unique negative correlations - Spearman	89,223	268,841	379
Unique negative correlations - Kendall	101,581	685,135	465

Table 5.5: Statistical comparison of different correlation methods, applied to the NOOB-normalised betas for Chromosome 21 (10300 sites)

For chromosome 21, we can see that there are a lot of consistent correlations (80% of the total), and that the Pearson correlation method produces the highest number of unique strong correlations. A similar trend can be seen with the sex chromosomes - the Pearson method consistently produces the highest number of unique strong correlations. There is a higher proportion of consistent strong correlations for chromosome 21 than the sex chromosomes.

5.4 Discussion

5.4.1 Computational considerations

The results in table 5.1 suggest that generation of the Pearson correlation method is the fastest, followed by the Spearman rank correlation matrix, then finally the Kendall rank correlation matrix which took a significantly longer time (orders of magnitude longer than the comparable Spearman method). From a performance point of view, calculating the Pearson correlation is fastest, though the Spearman method is typically only slower by less than an order of magnitude. Given the preference for a rank-correlation metric owing to the presumed non-linearity in the data (as discussed in section 1.3.2) the Spearman rank correlation appears to be the best choice.

It must be considered that profiling results can be significantly influenced by the efficiency of the implementation of whatever is being profiled. In our case, we are using the same Python library for all three methods; future studies may make use of a different software implementation which may improve performance for any of our tested methods so profiling results may vary.

5.4.2 Statistical considerations

The percentage of positive correlations was similar (within 1%) and mean correlation is within 0.1 for each method, for all tested chromosomes. This suggests that the underlying statistical distribution may be similar. This is supported by the high proportion of consistent correlations (>80% for all except positive correlations

on the Y chromosome) of the strongest 10% between each method for the tested chromosomes, shown in table 5.5.

In all cases, the mean correlation coefficient was significantly closer to zero for the Kendall matrices than those for the Spearman or Pearson methods. On its own, this would suggest a more 'centred' distribution (and thus a more even split between positive and negative correlations) but we can also see that the positive-to-negative ratio actually tends to be slightly greater than those of the other methods, so this doesn't appear to be the case. The main concern regarding distribution of correlation values and positive-to-negative ratio would be that a sufficiently-skewed distribution would make it difficult to identify a threshold (or proportion) for selecting strong correlations. All three methods produce a similar ratio of positive to negative correlations, though table 5.5 shows that there were some significant differences in which correlations were identified. The number of strong correlations that were found consistently regardless of correlation matrix was substantially higher than the number of strong correlations unique to each method, but there were still a large number of strong correlations unique to a particular method - we can use the numbers of unique strong correlations to identify the (likely) best method, following a similar logic to that discussed for normalisation types in section 4.6.3.2.

The Spearman approach produced the fewest unique strong correlations, which suggests that it performs better at capturing more 'legitimate' strong correlations, as more often not, its strongest correlations were backed up by at least one of the other methods - in other words, having fewer unique values means that it was less likely to pick up strong correlations that are only produced by the Spearman method. In the absence of evidence to the contrary, we must assume that a proportion of strong correlations aren't grounded in biology and only occur due to the mathematical nuances of the method used to generate them (if this weren't the case, there would be no correlations unique to each method). Assuming that this proportion is the same for all methods (a naïve approach, but we have no evidence to say otherwise) then the method that produces the fewest unique correlations is also the method that produces the highest number of legitimate correlations.

The Pearson method's strong correlations tended to have less overlap with either of the other methods. This is potentially due to the fact that Pearson correlation is less-suited to non-linear relationships, and we would expect underlying epigenetic phenomena to be highly non-linear in nature owing to the complexity of interactions happening at a molecular level. In this case, the high number of unique Pearson correlations show that there is certainly a difference in strong correlations detected via rank and non-rank methods, and we should err on the side of using rank methods as they are more suited for non-linear relationships.

Kendall's method generated correlation matrices with a lower overall variance than either of the other methods. Variance is a measure of spread that is particularly sensitive to outliers - in this case, the low variance suggests that there are fewer correlations at the extremes of the distribution than in other methods. This may make it difficult to select a threshold for a strong correlation if using the Kendall rank correlation coefficient. Conversely, Pearson's method had the highest overall variance in beta correlation matrix for all chromosomes tested, which would make it easier to find a threshold; however the assumption of linearity made for the Pearson coefficient means it is unlikely to be suitable for DNA methylation data.

5.5 Concluding remarks

In this chapter, we have investigated the effects of correlation method choice on the resulting correlation matrices produced from NOOB-normalised beta values, derived from several chromosomes, using the CHDS cohort. Our findings suggest that there are certainly differences in the outputs of each method, and this needs to be taken into account.

The Spearman rank correlation was the method of choice in the preliminary study of chapter 3, for reasons described in section 2.2.2. The results of studies in this chapter have supported the use of the Spearman method, for multiple reasons:

- Out of the options for a rank correlation coefficient (desired as the underlying data is assumed to be non-linear), the Spearman coefficient is orders of magnitude faster to compute than the Kendall coefficient, for the data that we tested, so we are able to assess correlations much faster.
- There are significant differences between rank and non-rank correlation methods, and given the presumed non-linearity of the underlying epigenetic associations, a rank correlation method would be preferable.

Further studies could validate results in this section using more chromosomes, different test cohorts, etc. For the time being, we consider there to be sufficient evidence to continue to use the Spearman rank correlation coefficient when evaluating correlations in methylation intensity.

Part III

Biological Studies

Chapter 6

Distance between correlating loci within a chromosome

6.1 Premise

One of our hypotheses is that CpG sites that are located closer together will tend to correlate more strongly in methylation intensity. In this chapter, we look at the distance between strongly-correlating loci on the same chromosome, and also how the typical strength of correlations change as this distance increases.

Studies have shown that functional gene groups are often co-located on the same chromosome (Thévenin et al. 2014). The physical 'architecture' of the genome is thought to play a role in gene interaction and regulation and recent techniques have begun to assess the functional implications of topologically-associating domains (Pombo and Dillon, 2015). Given the interplay between various different epigenetic mechanisms, it would be interesting to see if spatial associations in DNA methylation were also present. We would also expect to see some (relatively short-distance) associations owing to the presence of CpG islands (discussed in section 1.2) which would be likely to have a similar methylation state for many of its constituent CpG sites if it were associated with a promoter.

To test our hypothesis, four studies are undertaken:

- A comparison of correlation strength versus overall distance between sites located on the same chromosome
- A comparison of correlation strength versus overall distance between sites associated with the same gene (and chromosome)
- An analysis of where strongly-correlating loci are situated physically (within the same chromosome)
- An analysis of correlation within CpG islands

Discussion for the studies in this chapter is included in the general discussion (chapter 8).

6.2 Study: Correlation strength versus distance - per chromosome

In this study, we attempt to identify a relationship between correlation strength and distance between correlating loci. In the absence of any past studies suggesting the contrary, we assume that such a relationship would have a detectable linear component. A complete model for correlation strength probability versus distance between the correlating pair is beyond the scope of this thesis, though we may be able to identify linear components of a relationship via regression analysis.

Topologically-associating domains are extremely complex, reflecting the complexity of genome architecture as a whole, but the general idea is that there are regions in which compartments of the genome interact with each other more frequently than with regions outside that compartment (Pombo and Dillon, 2015). As chemical and physical properties arise for a given region of the genome due to the sequence of bases in that region, it follows that many of the self-interacting sequences would have to be at least partially contiguous; i.e. co-located within some relatively short stretch of the genome. The complexity of that interaction would certainly result in non-linear effects, but our simplified linear model may be capable of detecting some aspects of the overall trend.

6.2.1 Methods

Using the protocol specified in section 2.2, beta correlation matrices were calculated for a selection of chromosomes using data for the CHDS cohort (see appendix B) for the normalisation types selected on the basis of chapter 5:

- NOOB, for autosomes
- SWAN, for allosomes

Due to the computational intensity of this study, a selection of chromosomes were used, as per the rationale in 2.3.1. These chromosomes include: 9, 13, 15, 18, 20, 21, 22, X and Y.

The physical distance between each correlating pair was derived from Illumina's manifest by taking the difference between the position (on the chromosome) for each member of the pair. Linear regression analysis was performed between the correlation coefficient and the physical distance between the associated loci (using the *statsmodels* Python library), as this would indicate a relationship. Rather than taking the absolute magnitude of the correlation and generating a single regression model, positive and negative correlations were regressed separately as studies in chapter 4 suggested that they have different statistical properties, so combining them may result in less-accurate models.

6.2.2 Results

	Positive Correlations	Negative Correlations
Number of observations	210811274	131531587
R-Squared	0.0	0.0
F-score	7880.0	52.8
F-score P-value	0.0	3.72e-13
Gradient	-1.78e-11	1.36e-12
Gradient standard error	2e-13	1.87e-13
Gradient T-score	-88.7	7.26
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.82e-11, -1.74e-11)	(9.92e-13, 1.72e-12)
Intercept	0.134	-0.0954
Intercept standard error	1.24e-05	1.16e-05
Intercept T-score	10800.0	-8210.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.134, 0.134)	(-0.095, -0.095)

Table 6.1: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 9, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	136109854	85220426
R-Squared	0.0	0.0
F-score	60700.0	11400.0
F-score P-value	0.0	0.0
Gradient	-9.15e-11	3.63e-11
Gradient standard error	3.71e-13	3.4e-13
Gradient T-score	-246.0	107.0
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-9.22e-11, -9.08e-11)	(3.57e-11, 3.7e-11)
Intercept	0.146	-0.101
Intercept standard error	1.8e-05	1.64e-05
Intercept T-score	8090.0	-6150.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.146, 0.146)	(-0.101, -0.101)

Table 6.2: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 13, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	254297067	158711103
R-Squared	0.0	0.0
F-score	91100.0	27000.0
F-score P-value	0.0	0.0
Gradient	-1.18e-10	6.13e-11
Gradient standard error	3.92e-13	3.73e-13
Gradient T-score	-302.0	164.0
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.19e-10, -1.18e-10)	(6.05e-11, 6.2e-11)
Intercept	0.138	-0.0987
Intercept standard error	1.26e-05	1.19e-05
Intercept T-score	10900.0	-8290.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.138, 0.138)	(-0.099, -0.099)

Table 6.3: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 15, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	68742921	42239730
R-Squared	0.0	0.0
F-score	17200.0	1850.0
F-score P-value	0.0	0.0
Gradient	-9.78e-11	2.9e-11
Gradient standard error	7.46e-13	6.74e-13
Gradient T-score	-131.0	43.0
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-9.93e-11, -9.64e-11)	(2.77e-11, 3.03e-11)
Intercept	0.146	-0.1
Intercept standard error	2.59e-05	2.35e-05
Intercept T-score	5630.0	-4270.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.146, 0.146)	(-0.1, -0.1)

Table 6.4: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 18, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	165890096	97679224
R-Squared	0.0	0.0
F-score	1620.0	1420.0
F-score P-value	0.0	4.87e-310
Gradient	-2.15e-11	-1.97e-11
Gradient standard error	5.33e-13	5.23e-13
Gradient T-score	-40.3	-37.6
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-2.25e-11, -2.04e-11)	(-2.07e-11, -1.87e-11)
Intercept	0.131	-0.0931
Intercept standard error	1.45e-05	1.44e-05
Intercept T-score	9040.0	-6470.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.131, 0.131)	(-0.093, -0.093)

Table 6.5: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 20, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	33906857	19132993
R-Squared	0.0	0.0
F-score	2180.0	1560.0
F-score P-value	0.0	0.0
Gradient	1.12e-10	9.76e-11
Gradient standard error	2.39e-12	2.47e-12
Gradient T-score	46.7	39.4
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(1.07e-10, 1.16e-10)	(9.27e-11, 1.02e-10)
Intercept	0.136	-0.0961
Intercept standard error	3.05e-05	3.11e-05
Intercept T-score	4450.0	-3090.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.136, 0.136)	(-0.096, -0.096)

Table 6.6: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 21, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	108741296	59922865
R-Squared	0.0	0.0
F-score	9480.0	3780.0
F-score P-value	0.0	0.0
Gradient	-1.19e-10	8.04e-11
Gradient standard error	1.22e-12	1.31e-12
Gradient T-score	-97.3	61.5
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.21e-10, -1.16e-10)	(7.78e-11, 8.29e-11)
Intercept	0.131	-0.0931
Intercept standard error	1.77e-05	1.89e-05
Intercept T-score	7420.0	-4920.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.131, 0.131)	(-0.093, -0.093)

Table 6.7: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome 22, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	97951388	84253117
R-Squared	0.001	0.001
F-score	142000.0	65000.0
F-score P-value	0.0	0.0
Gradient	-2.09e-10	1.3e-10
Gradient standard error	5.56e-13	5.08e-13
Gradient T-score	-376.0	255.0
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-2.1e-10, -2.08e-10)	(1.29e-10, 1.31e-10)
Intercept	0.372	-0.329
Intercept standard error	3.76e-05	3.42e-05
Intercept T-score	9910.0	-9610.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.372, 0.373)	(-0.329, -0.329)

Table 6.8: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome x, using swan-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	83330	60586
R-Squared	0.01	0.001
F-score	826.0	82.8
F-score P-value	1.15e-180	9.41e-20
Gradient	-2.89e-09	9.71e-10
Gradient standard error	1e-10	1.07e-10
Gradient T-score	-28.7	9.1
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-3.08e-09, -2.69e-09)	(7.62e-10, 1.18e-09)
Intercept	0.465	-0.379
Intercept standard error	0.001	0.001
Intercept T-score	448.0	-338.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.463, 0.467)	(-0.382, -0.377)

Table 6.9: Linear regression results for strength of correlation versus distance between correlating pair, for chromosome y, using swan-normalised data (CHDS dataset).

The results show that the R-squared values for our linear models is at or near zero in all cases. R-squared can be thought of as a measure of the predictive capability of a model - as all of our models have such a low R-squared, the predictive capability of our linear regression model for all tested chromosomes is practically negligible; in other words, the vast majority of the variation in the data cannot be explained by our regression line. Nonetheless, details for each of the models are retained in this section for posterity.

6.3 Study: Correlation strength versus distance - within genes

In section 6.2, we investigated the general trend of correlation strength versus distance with chromosomes. In this study, we use a similar approach and justification to investigate the association between correlation strength and distance for CpG sites within the same genes. One of our overarching hypotheses is that CpG sites located close to each other will tend to correlate strongly in methylation intensity; we can take this further and postulate that CpG sites in the same functional group will tend to demonstrate similar effects.

6.3.1 Methods

For this study, the methods described in section 6.2.1 were adapted such that they only looked at correlations that were on the same gene. The majority of the process is the same, with the main difference being that correlations were only included in the regression analysis if their constituent CpG sites are on the same gene. This difference also significantly reduces the computational intensity of the study, so we can look at all chromosomes rather than a subset. An additional linear regression was performed on all same-gene correlating pairs from all autosomal chromosomes.

6.3.2 Results

	Positive Correlations	Negative Correlations
Number of observations	1128773	579331
R-Squared	0.0	0.0
F-score	3.18	0.00033
F-score P-value	0.0744	0.986
Gradient	-6.79e-11	-5.71e-13
Gradient standard error	3.81e-11	3.14e-11
Gradient T-score	-1.78	-0.018
Gradient $P > \|T\|$	0.074	0.986
Gradient 95% confidence interval	(-1.43e-10, 6.69e-12)	(-6.22e-11, 6.1e-11)
Intercept	0.161	-0.102
Intercept standard error	0.0	0.0
Intercept T-score	1220.0	-821.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.161, 0.161)	(-0.102, -0.101)

Table 6.10: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 1, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	790649	424188
R-Squared	0.002	0.0
F-score	1660.0	23.8
F-score P-value	0.0	1.05e-06
Gradient	-4.86e-08	4.93e-09
Gradient standard error	1.19e-09	1.01e-09
Gradient T-score	-40.7	4.88
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-5.09e-08, -4.62e-08)	(2.95e-09, 6.91e-09)
Intercept	0.171	-0.104
Intercept standard error	0.0	0.0
Intercept T-score	837.0	-572.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.17, 0.171)	(-0.104, -0.104)

Table 6.11: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 2, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	554420	322046
R-Squared	0.0	0.002
F-score	25.7	580.0
F-score P-value	4.05e-07	4.68e-128
Gradient	-6.8e-09	-2.56e-08
Gradient standard error	1.34e-09	1.06e-09
Gradient T-score	-5.07	-24.1
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-9.43e-09, -4.17e-09)	(-2.77e-08, -2.35e-08)
Intercept	0.177	-0.108
Intercept standard error	0.0	0.0
Intercept T-score	691.0	-485.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.176, 0.177)	(-0.108, -0.108)

Table 6.12: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 3, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	461036	248332
R-Squared	0.002	0.0
F-score	1040.0	32.4
F-score P-value	1.74e-228	1.24e-08
Gradient	-5.47e-08	-7.64e-09
Gradient standard error	1.69e-09	1.34e-09
Gradient T-score	-32.3	-5.69
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-5.8e-08, -5.14e-08)	(-1.03e-08, -5.01e-09)
Intercept	0.17	-0.102
Intercept standard error	0.0	0.0
Intercept T-score	642.0	-445.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.17, 0.171)	(-0.102, -0.101)

Table 6.13: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 4, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	538688	284307
R-Squared	0.0	0.001
F-score	1.61	212.0
F-score P-value	0.204	5.52e-48
Gradient	-1.63e-09	-1.52e-08
Gradient standard error	1.28e-09	1.04e-09
Gradient T-score	-1.27	-14.6
Gradient $P > \|T\|$	0.204	0.0
Gradient 95% confidence interval	(-4.14e-09, 8.84e-10)	(-1.72e-08, -1.31e-08)
Intercept	0.18	-0.106
Intercept standard error	0.0	0.0
Intercept T-score	732.0	-487.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.18, 0.181)	(-0.106, -0.106)

Table 6.14: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 5, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	795144	435094
R-Squared	0.0	0.0
F-score	0.163	21.8
F-score P-value	0.686	2.97e-06
Gradient	-1.13e-10	-1.23e-09
Gradient standard error	2.79e-10	2.63e-10
Gradient T-score	-0.404	-4.67
Gradient $P > \|T\|$	0.686	0.0
Gradient 95% confidence interval	(-6.59e-10, 4.34e-10)	(-1.74e-09, -7.13e-10)
Intercept	0.166	-0.104
Intercept standard error	0.0	0.0
Intercept T-score	994.0	-708.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.165, 0.166)	(-0.104, -0.104)

Table 6.15: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 6, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	1669987	809132
R-Squared	0.006	0.002
F-score	10400.0	1550.0
F-score P-value	0.0	0.0
Gradient	-3.74e-08	1.37e-08
Gradient standard error	3.67e-10	3.48e-10
Gradient T-score	-102.0	39.4
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-3.81e-08, -3.67e-08)	(1.3e-08, 1.44e-08)
Intercept	0.151	-0.0933
Intercept standard error	0.0	0.0
Intercept T-score	1200.0	-754.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.15, 0.151)	(-0.094, -0.093)

Table 6.16: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 7, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	504443	274149
R-Squared	0.0	0.0
F-score	241.0	7.13
F-score P-value	2.42e-54	0.00757
Gradient	-9.61e-09	1.64e-09
Gradient standard error	6.19e-10	6.13e-10
Gradient T-score	-15.5	2.67
Gradient $P > \|T\|$	0.0	0.008
Gradient 95% confidence interval	(-1.08e-08, -8.4e-09)	(4.36e-10, 2.84e-09)
Intercept	0.163	-0.102
Intercept standard error	0.0	0.0
Intercept T-score	764.0	-528.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.163, 0.164)	(-0.103, -0.102)

Table 6.17: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 8, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	256050	134440
R-Squared	0.001	0.0
F-score	153.0	43.3
F-score P-value	3.16e-35	4.75e-11
Gradient	-2.46e-08	-1.18e-08
Gradient standard error	1.99e-09	1.8e-09
Gradient T-score	-12.4	-6.58
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-2.85e-08, -2.07e-08)	(-1.53e-08, -8.29e-09)
Intercept	0.165	-0.1
Intercept standard error	0.0	0.0
Intercept T-score	518.0	-343.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.164, 0.166)	(-0.101, -0.099)

Table 6.18: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 9, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	829795	419801
R-Squared	0.002	0.0
F-score	1600.0	68.1
F-score P-value	0.0	1.54e-16
Gradient	-4.42e-08	8.24e-09
Gradient standard error	1.1e-09	9.98e-10
Gradient T-score	-40.0	8.25
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-4.63e-08, -4.2e-08)	(6.28e-09, 1.02e-08)
Intercept	0.163	-0.101
Intercept standard error	0.0	0.0
Intercept T-score	836.0	-538.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.163, 0.164)	(-0.101, -0.1)

Table 6.19: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 10, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	731467	364664
R-Squared	0.002	0.0
F-score	1470.0	133.0
F-score P-value	4.6e-321	1.05e-30
Gradient	-3.3e-08	9.79e-09
Gradient standard error	8.61e-10	8.5e-10
Gradient T-score	-38.3	11.5
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-3.47e-08, -3.13e-08)	(8.13e-09, 1.15e-08)
Intercept	0.163	-0.102
Intercept standard error	0.0	0.0
Intercept T-score	861.0	-558.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.162, 0.163)	(-0.102, -0.101)

Table 6.20: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 11, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	533400	281389
R-Squared	0.0	0.0
F-score	24.1	3.52
F-score P-value	9.35e-07	0.0606
Gradient	-4.05e-09	-2.59e-09
Gradient standard error	8.25e-10	1.38e-09
Gradient T-score	-4.91	-1.88
Gradient $P > \|T\|$	0.0	0.061
Gradient 95% confidence interval	(-5.67e-09, -2.43e-09)	(-5.29e-09, 1.15e-10)
Intercept	0.165	-0.101
Intercept standard error	0.0	0.0
Intercept T-score	777.0	-485.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.165, 0.166)	(-0.102, -0.101)

Table 6.21: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 12, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	359881	176484
R-Squared	0.0	0.001
F-score	3.41	130.0
F-score P-value	0.0648	3.64e-30
Gradient	-3.27e-09	-1.84e-08
Gradient standard error	1.77e-09	1.61e-09
Gradient T-score	-1.85	-11.4
Gradient $P > \|T\|$	0.065	0.0
Gradient 95% confidence interval	(-6.73e-09, 2.01e-10)	(-2.16e-08, -1.52e-08)
Intercept	0.161	-0.0988
Intercept standard error	0.0	0.0
Intercept T-score	579.0	-358.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.16, 0.161)	(-0.099, -0.098)

Table 6.22: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 13, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	313867	168714
R-Squared	0.0	0.0
F-score	76.3	18.6
F-score P-value	2.47e-18	1.6e-05
Gradient	-1.07e-08	-4.49e-09
Gradient standard error	1.22e-09	1.04e-09
Gradient T-score	-8.73	-4.31
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.31e-08, -8.3e-09)	(-6.54e-09, -2.45e-09)
Intercept	0.168	-0.105
Intercept standard error	0.0	0.0
Intercept T-score	586.0	-400.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.168, 0.169)	(-0.106, -0.105)

Table 6.23: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 14, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	341288	179510
R-Squared	0.002	0.0
F-score	809.0	42.2
F-score P-value	1.02e-177	8.3e-11
Gradient	-6.67e-08	-1.32e-08
Gradient standard error	2.35e-09	2.03e-09
Gradient T-score	-28.4	-6.5
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-7.13e-08, -6.21e-08)	(-1.72e-08, -9.2e-09)
Intercept	0.169	-0.101
Intercept standard error	0.0	0.0
Intercept T-score	555.0	-362.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.168, 0.169)	(-0.102, -0.1)

Table 6.24: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 15, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	562970	275649
R-Squared	0.001	0.0
F-score	399.0	17.0
F-score P-value	1.14e-88	3.66e-05
Gradient	-2.27e-08	-4.55e-09
Gradient standard error	1.13e-09	1.1e-09
Gradient T-score	-20.0	-4.13
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-2.49e-08, -2.04e-08)	(-6.71e-09, -2.39e-09)
Intercept	0.154	-0.0978
Intercept standard error	0.0	0.0
Intercept T-score	788.0	-492.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.154, 0.155)	(-0.098, -0.097)

Table 6.25: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 16, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	750461	371595
R-Squared	0.005	0.0
F-score	3770.0	67.8
F-score P-value	0.0	1.83e-16
Gradient	-1.09e-07	-1.38e-08
Gradient standard error	1.77e-09	1.67e-09
Gradient T-score	-61.4	-8.23
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.12e-07, -1.05e-07)	(-1.7e-08, -1.05e-08)
Intercept	0.16	-0.0978
Intercept standard error	0.0	0.0
Intercept T-score	821.0	-491.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.16, 0.161)	(-0.098, -0.097)

Table 6.26: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 17, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	183691	96735
R-Squared	0.0	0.001
F-score	40.2	106.0
F-score P-value	2.24e-10	7.43e-25
Gradient	-1.97e-08	-2.7e-08
Gradient standard error	3.11e-09	2.62e-09
Gradient T-score	-6.34	-10.3
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-2.58e-08, -1.36e-08)	(-3.21e-08, -2.18e-08)
Intercept	0.174	-0.103
Intercept standard error	0.0	0.0
Intercept T-score	380.0	-249.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.173, 0.175)	(-0.104, -0.102)

Table 6.27: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 18, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	332414	169477
R-Squared	0.009	0.0
F-score	3150.0	44.4
F-score P-value	0.0	2.63e-11
Gradient	-4.03e-07	-4.17e-08
Gradient standard error	7.18e-09	6.26e-09
Gradient T-score	-56.1	-6.67
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-4.17e-07, -3.89e-07)	(-5.4e-08, -2.95e-08)
Intercept	0.161	-0.0962
Intercept standard error	0.0	0.0
Intercept T-score	555.0	-355.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.16, 0.161)	(-0.097, -0.096)

Table 6.28: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 19, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	269591	124097
R-Squared	0.002	0.001
F-score	659.0	126.0
F-score P-value	2.97e-145	3.59e-29
Gradient	-4.13e-08	-1.37e-08
Gradient standard error	1.61e-09	1.22e-09
Gradient T-score	-25.7	-11.2
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-4.44e-08, -3.81e-08)	(-1.61e-08, -1.13e-08)
Intercept	0.167	-0.0945
Intercept standard error	0.0	0.0
Intercept T-score	530.0	-338.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.166, 0.167)	(-0.095, -0.094)

Table 6.29: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 20, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	114552	58678
R-Squared	0.001	0.002
F-score	93.8	94.4
F-score P-value	3.62e-22	2.73e-22
Gradient	-5.08e-08	-4.42e-08
Gradient standard error	5.25e-09	4.55e-09
Gradient T-score	-9.68	-9.71
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-6.11e-08, -4.05e-08)	(-5.32e-08, -3.53e-08)
Intercept	0.168	-0.1
Intercept standard error	0.001	0.0
Intercept T-score	325.0	-207.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.167, 0.169)	(-0.101, -0.099)

Table 6.30: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 21, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	223608	103282
R-Squared	0.004	0.0
F-score	884.0	23.6
F-score P-value	8.33e-194	1.17e-06
Gradient	-9.85e-08	-1.53e-08
Gradient standard error	3.31e-09	3.15e-09
Gradient T-score	-29.7	-4.86
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.05e-07, -9.2e-08)	(-2.15e-08, -9.14e-09)
Intercept	0.166	-0.0971
Intercept standard error	0.0	0.0
Intercept T-score	457.0	-271.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.165, 0.166)	(-0.098, -0.096)

Table 6.31: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome 22, using noob-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	114812	72079
R-Squared	0.021	0.002
F-score	2440.0	165.0
F-score P-value	0.0	1.06e-37
Gradient	-1.26e-07	3.57e-08
Gradient standard error	2.55e-09	2.78e-09
Gradient T-score	-49.3	12.8
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.31e-07, -1.21e-07)	(3.02e-08, 4.11e-08)
Intercept	0.422	-0.324
Intercept standard error	0.001	0.001
Intercept T-score	602.0	-428.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.42, 0.423)	(-0.326, -0.323)

Table 6.32: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome x, using swan-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	1124	622
R-Squared	0.022	0.089
F-score	25.4	60.4
F-score P-value	5.45e-07	3.2e-14
Gradient	-2.48e-07	4.65e-07
Gradient standard error	4.91e-08	5.98e-08
Gradient T-score	-5.04	7.77
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-3.44e-07, -1.51e-07)	(3.47e-07, 5.82e-07)
Intercept	0.543	-0.429
Intercept standard error	0.006	0.008
Intercept T-score	86.0	-53.3
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.531, 0.556)	(-0.445, -0.413)

Table 6.33: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, for chromosome y, using swan-normalised data (CHDS dataset).

	Positive Correlations	Negative Correlations
Number of observations	12246175	6301094
R-Squared	0.0	0.0
F-score	727.0	14.0
F-score P-value	4.65e-160	0.000182
Gradient	-1.01e-09	1.15e-10
Gradient standard error	3.74e-11	3.08e-11
Gradient T-score	-27.0	3.74
Gradient $P > \|T\|$	0.0	0.0
Gradient 95% confidence interval	(-1.08e-09, -9.35e-10)	(5.48e-11, 1.75e-10)
Intercept	0.16	-0.101
Intercept standard error	4.06e-05	3.75e-05
Intercept T-score	3950.0	-2680.0
Intercept $P > \|T\|$	0.0	0.0
Intercept 95% confidence interval	(0.16, 0.16)	(-0.101, -0.101)

Table 6.34: Linear regression results for strength of correlation versus distance between correlating pairs within the same gene, across all chromosomes, using noob-normalised data (CHDS dataset).

We saw in section 6.2.1 that our linear models had very limited predictive capability, owing to their low R-squared values. The same can be seen in our results here, for the most part. The sex chromosomes are the notable exceptions, with a substantially-higher R-squared than all autosomal models. With models for positive correlations having an R-squared of around 0.02 in both cases, and the negative correlation model for chromosome Y having an R-squared of almost 0.09, predictive power is still extremely low, though we can see some semblance of a trend emerging here. As we've done previously, details for each of the models are retained for posterity.

6.4 Study: Distance between strongly-correlating loci

Previously in this chapter, we have used linear regression to test our hypothesis that CpG sites located closer together will tend to correlate more strongly in methylation intensity. In this study, we take a different approach - rather than attempt to generate a model relating distance to correlation for some subset of CpG sites within the genome, we instead compare the distance graphically. Through this method, we can identify notable clusters of highly-correlating pairs in close proximity, which may help us identify parts of the genome that tend to associate epigenetically.

6.4.1 Methods

Raw data was processed and correlations generated as per the methods discussed in section 2.2, for the CHDS cohort. We took a proportional threshold approach (section 2.2.5.2), selecting the strongest 10% positive and weakest correlations as the strongest. Due to proportional thresholds requiring significant computational resource for large chromosomes, we only use this method to look at a subset of chromosomes, as per section 2.3.1. The subset of chromosomes used for this study includes: 4, 8, 9, 13, 14, 15, 16, 18, 20, 21, 22, X and Y.

One of the key difficulties is that a strong correlation may only be present if a specific normalisation type is used prior to correlation calculation. As a solution to this, we apply the *metaN* method as described in section 4.6.4.1 - that is, we consider a correlation to be more meaningful if it can be identified in datasets produced by multiple normalisation methods (including no normalisation).

To display the distances between these correlations, we plot the chromosomal positions of the two CpG sites on the X and Y axis respectively. This allows us to get a general idea of where the strong correlations are occuring, and identify areas that are particularly dense in strong correlations. Correlations are plotted in green if they are positive, and red if they are negative. If for some reason they are present in both (i.e. a pair is simultaneously strong and positive for one method, but strong and negative for another), they are plotted in black, though this is not considered to be a likely occurrence. Transparency of a given point on the graph is dependent on the number of times it was identified as a strong correlation in a dataset produced by one of the six methods of normalisation. Strongly-correlating pairs consistent across all six datasets are plotted with a darker shade of red or green.

6.4.2 Results

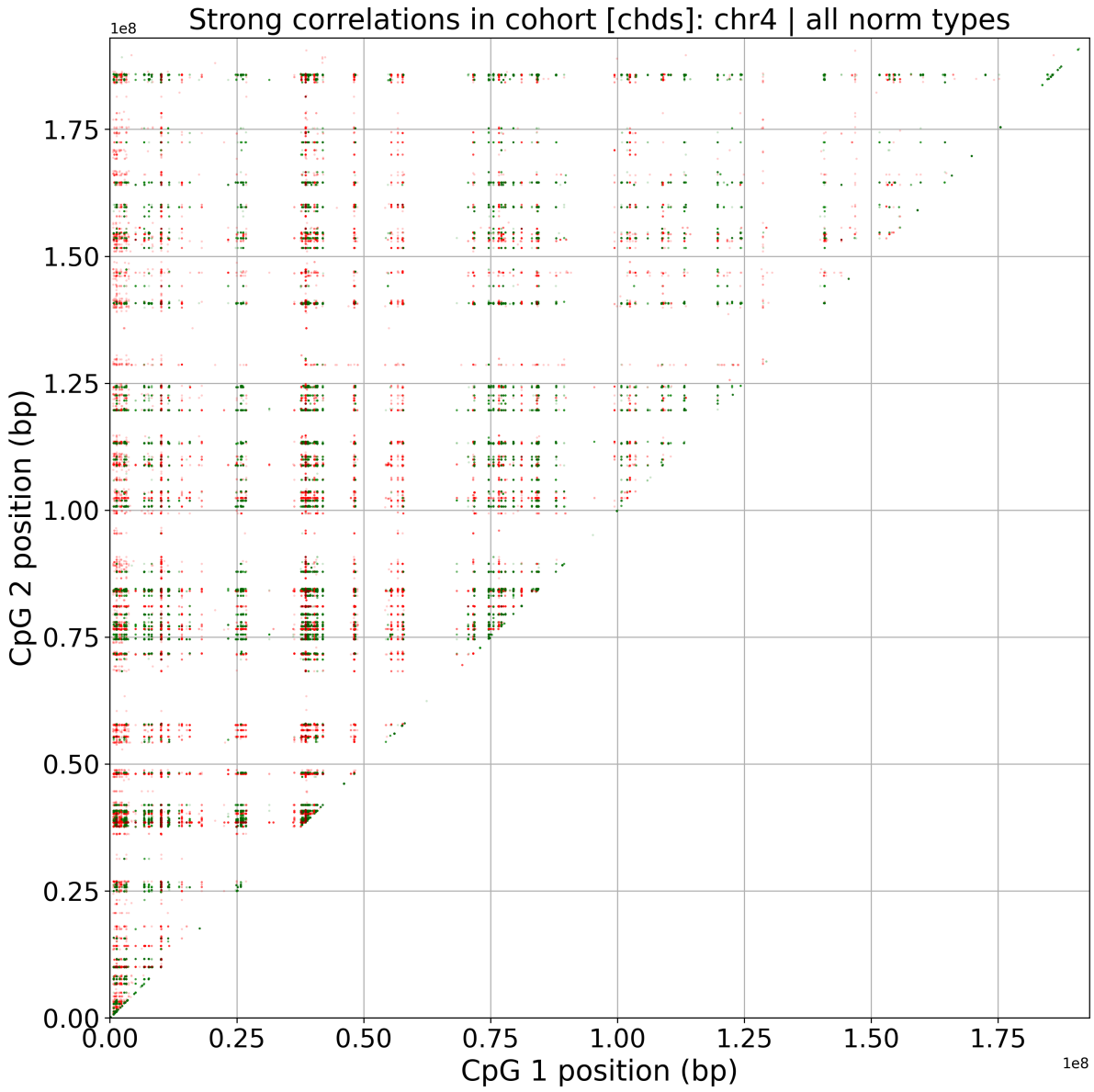

Figure 6.1: Location of strong correlations in CpG methylation intensity (beta) for chromosome 4, with intensity based on score for the metaN method. Source data: CHDS dataset

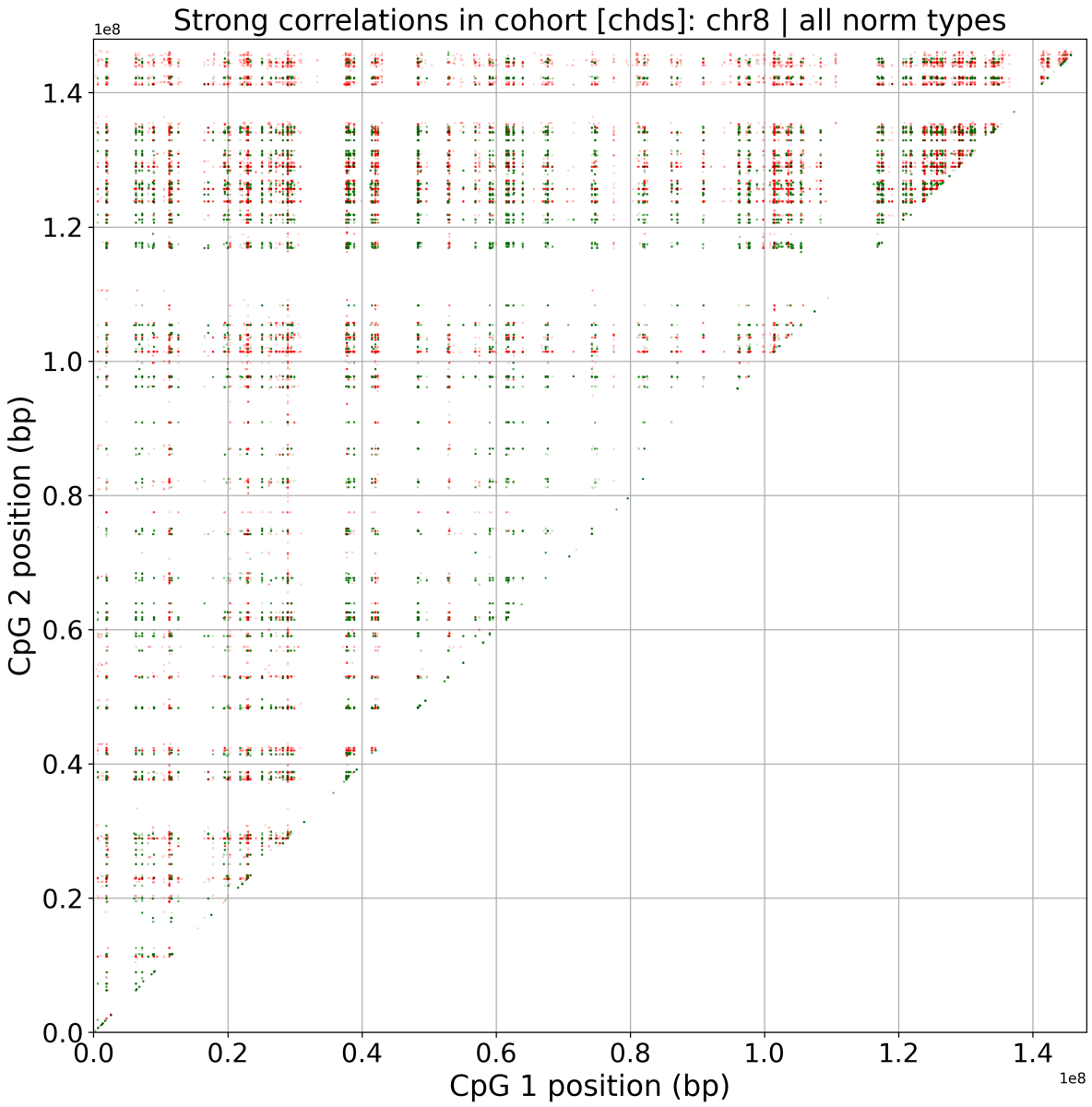

Figure 6.2: Location of strong correlations in CpG methylation intensity (beta) for chromosome 8, with intensity based on score for the metaN method. Source data: CHDS dataset

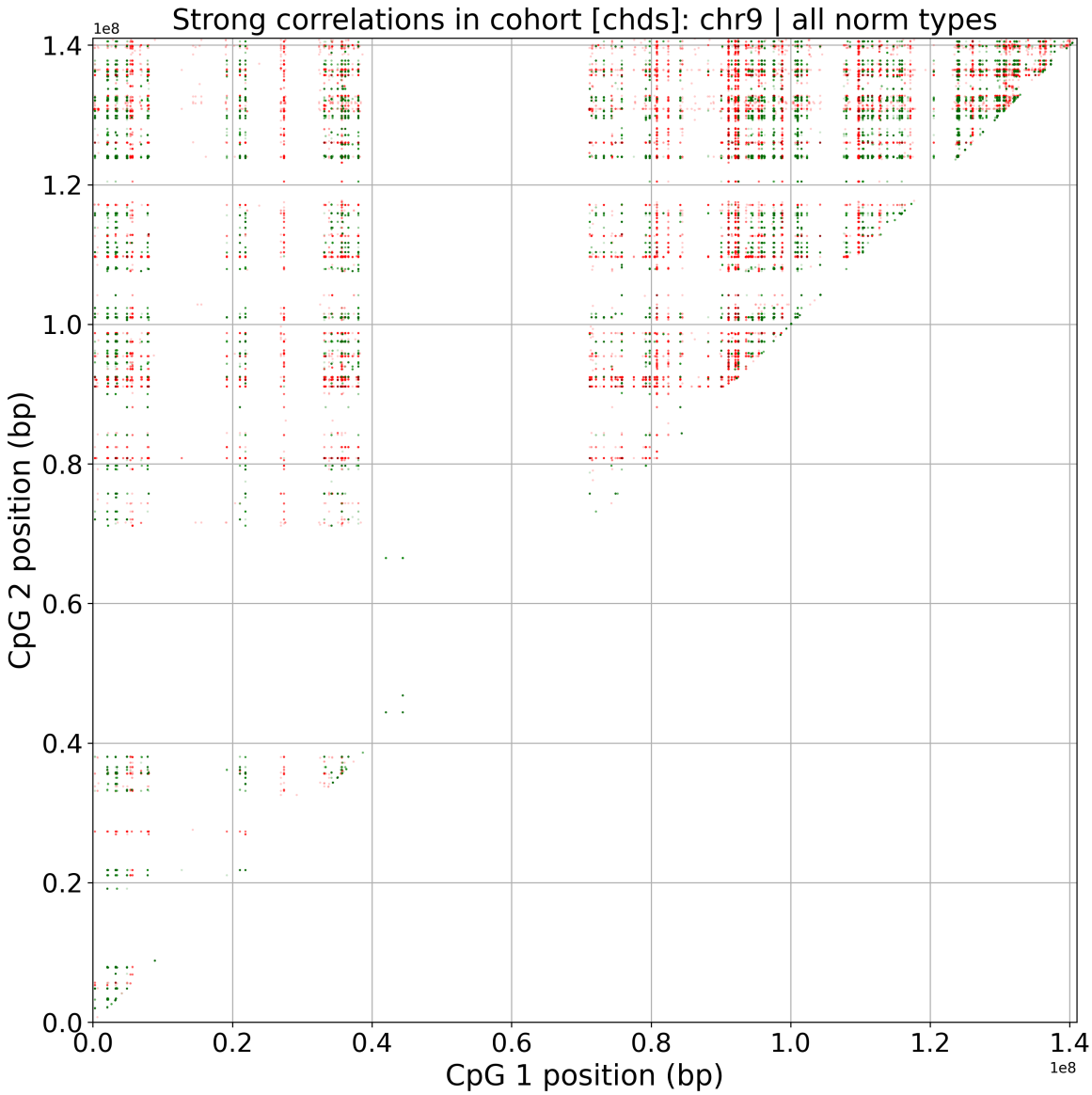

Figure 6.3: Location of strong correlations in CpG methylation intensity (beta) for chromosome 9, with intensity based on score for the metaN method. Source data: CHDS dataset

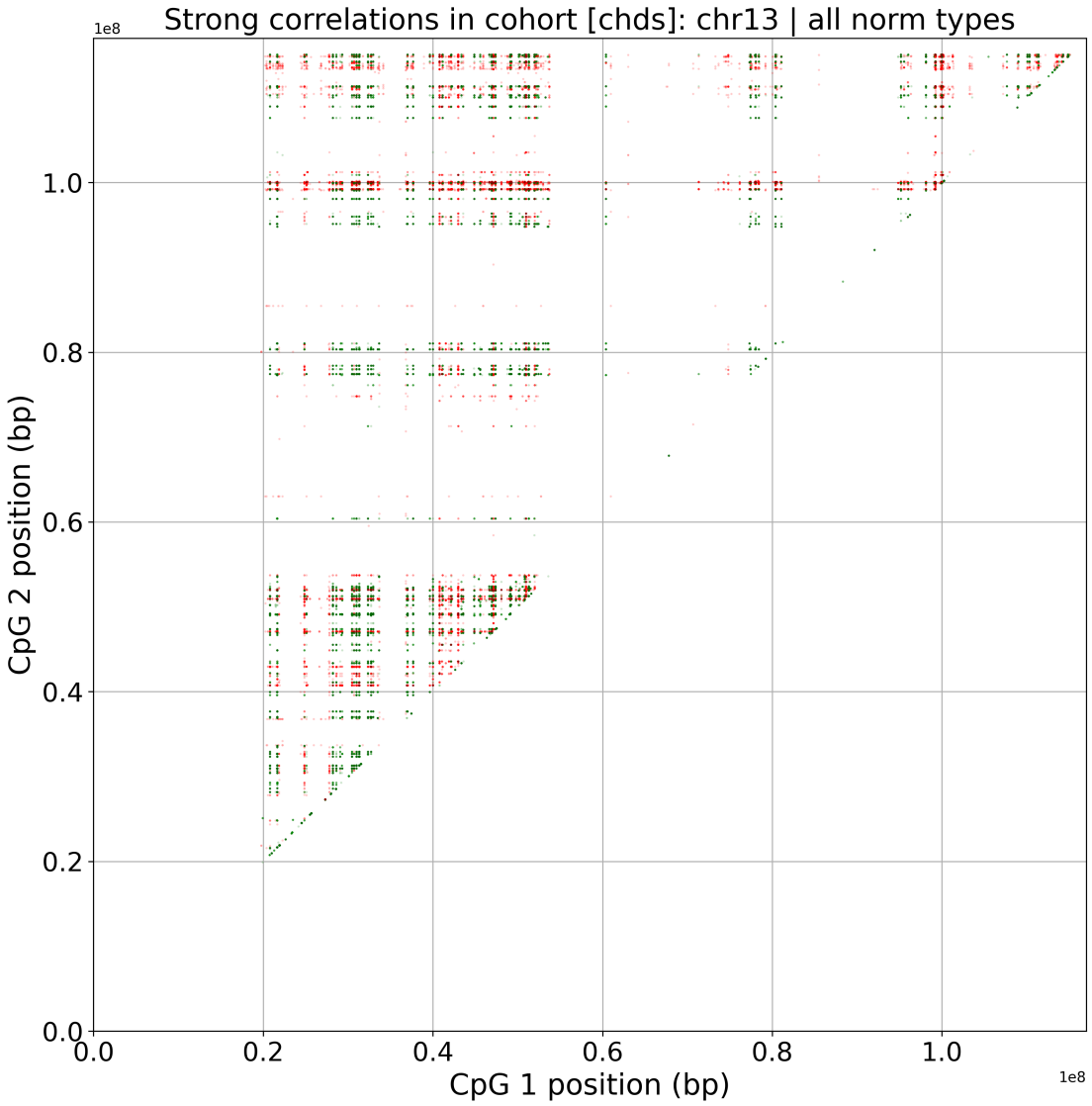

Figure 6.4: Location of strong correlations in CpG methylation intensity (beta) for chromosome 13, with intensity based on score for the metaN method. Source data: CHDS dataset

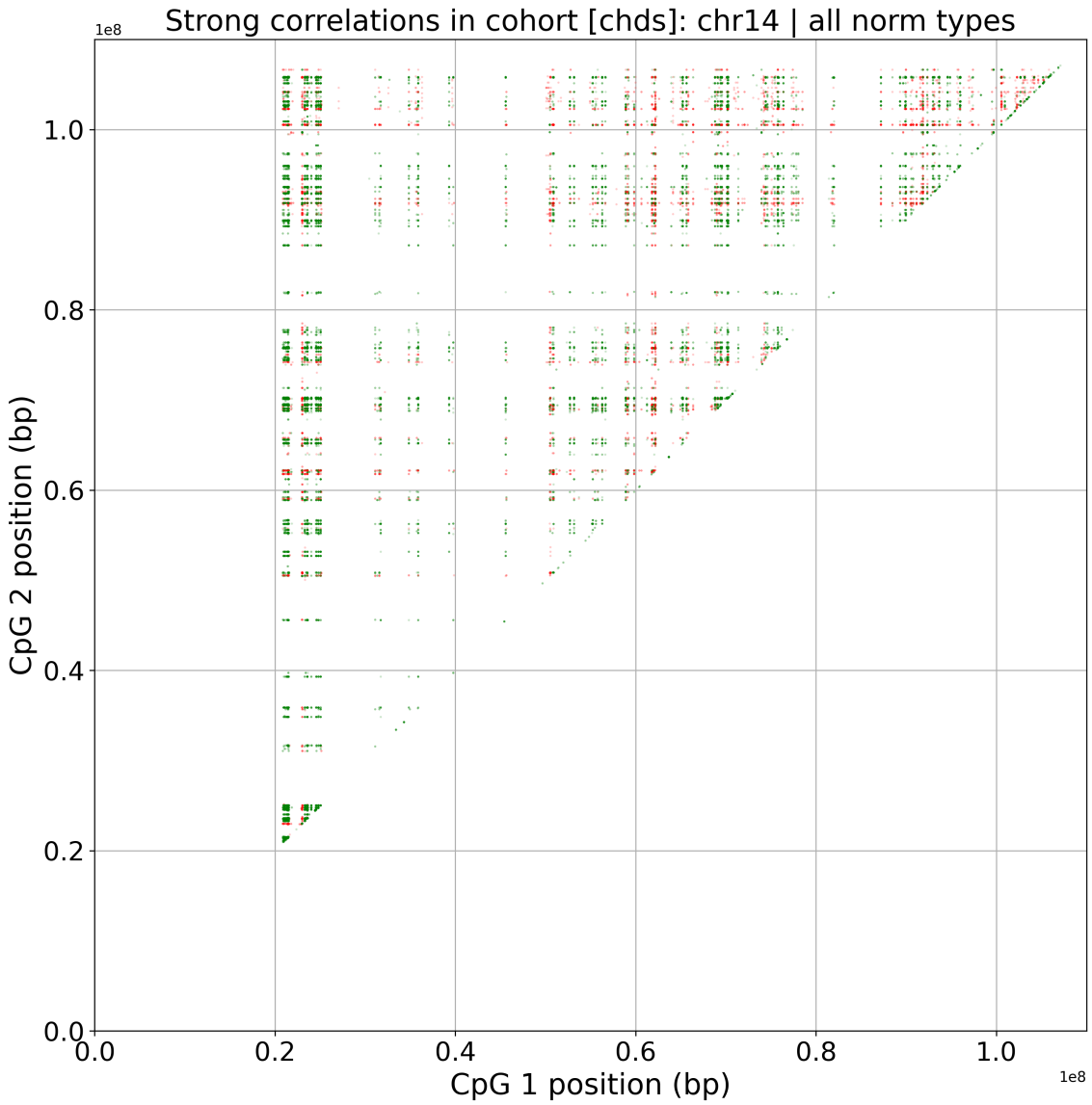

Figure 6.5: Location of strong correlations in CpG methylation intensity (beta) for chromosome 14, with intensity based on score for the metaN method. Source data: CHDS dataset

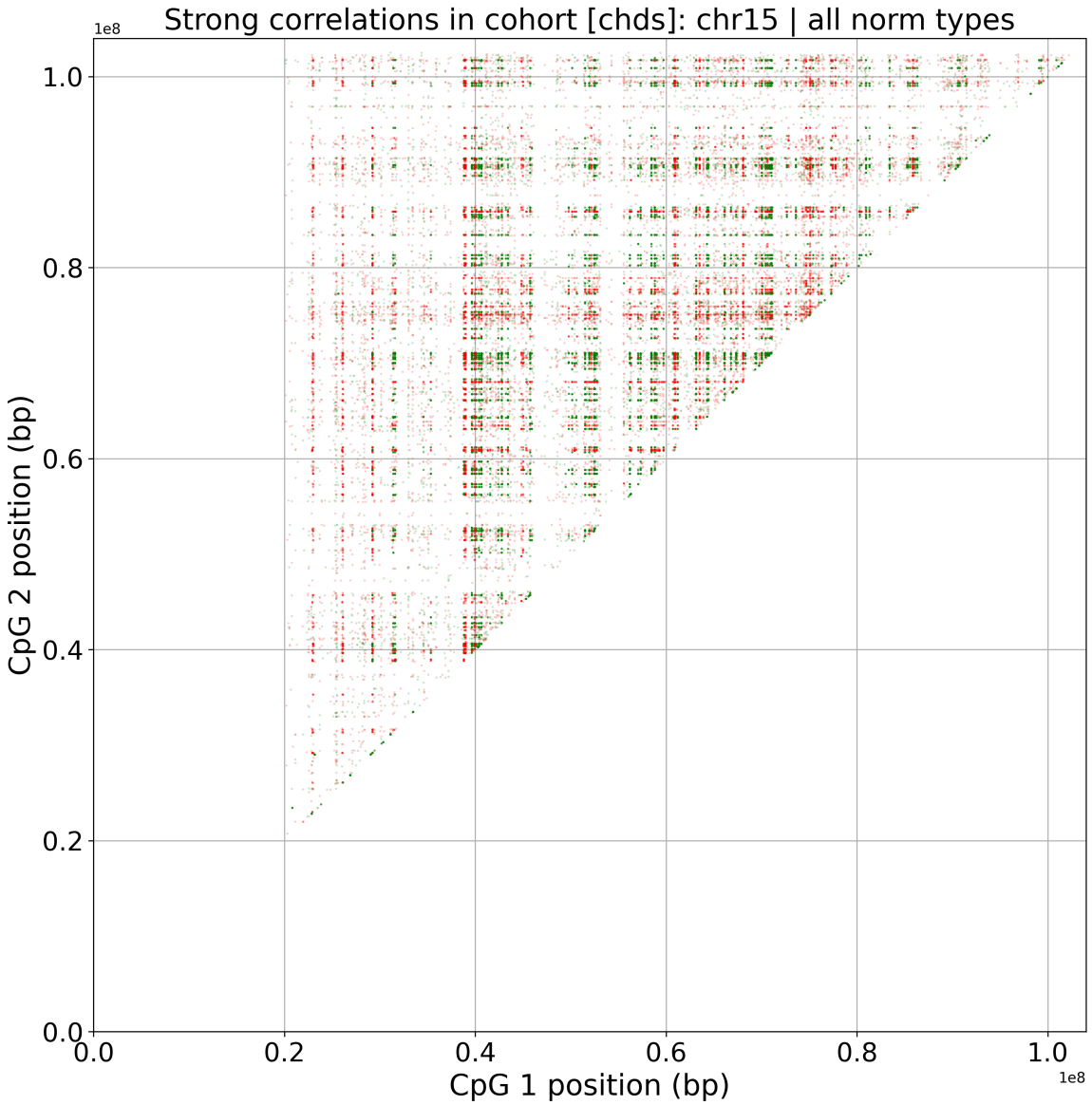

Figure 6.6: Location of strong correlations in CpG methylation intensity (beta) for chromosome 15, with intensity based on score for the metaN method. Source data: CHDS dataset

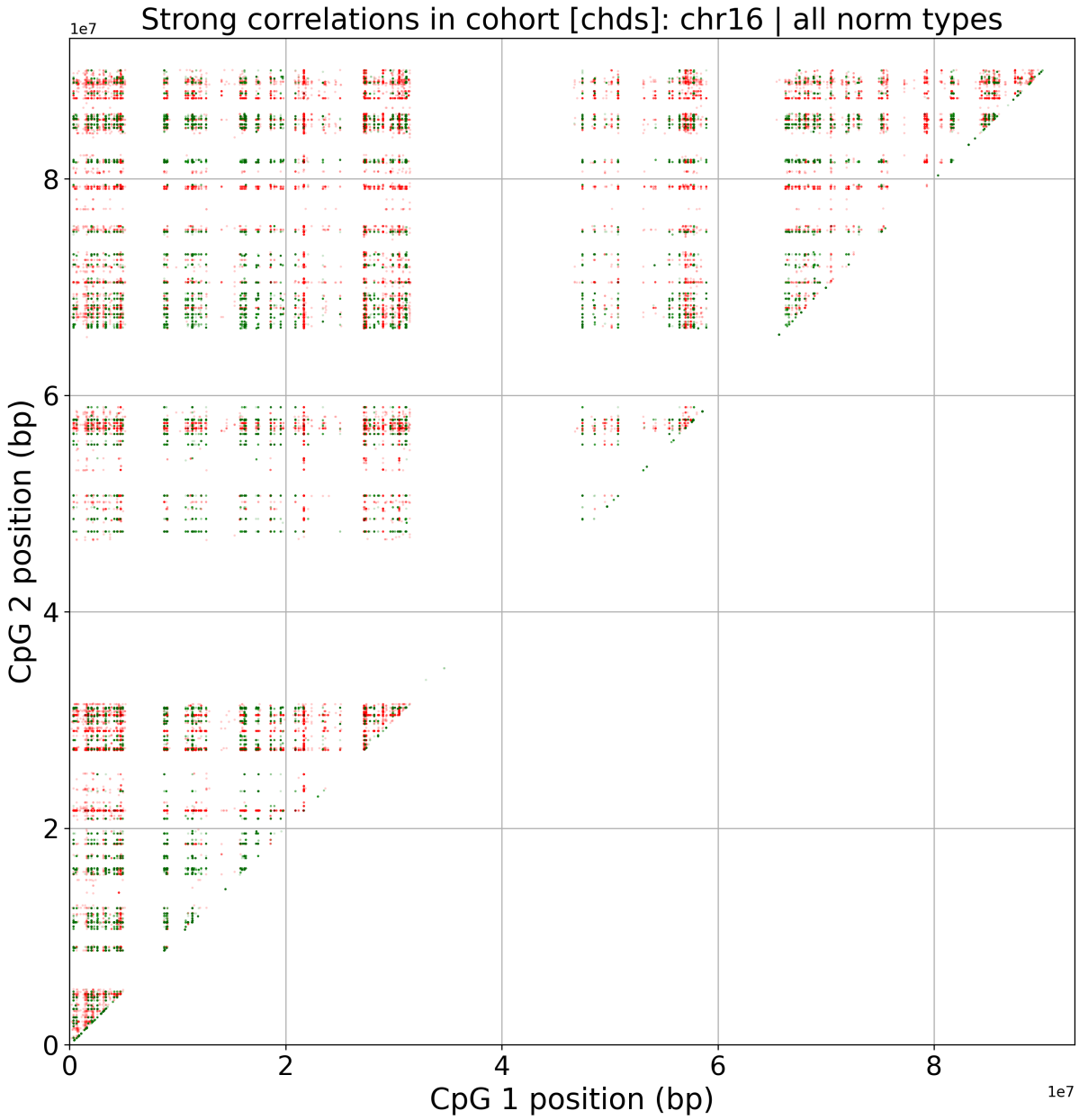

Figure 6.7: Location of strong correlations in CpG methylation intensity (beta) for chromosome 16, with intensity based on score for the metaN method. Source data: CHDS dataset

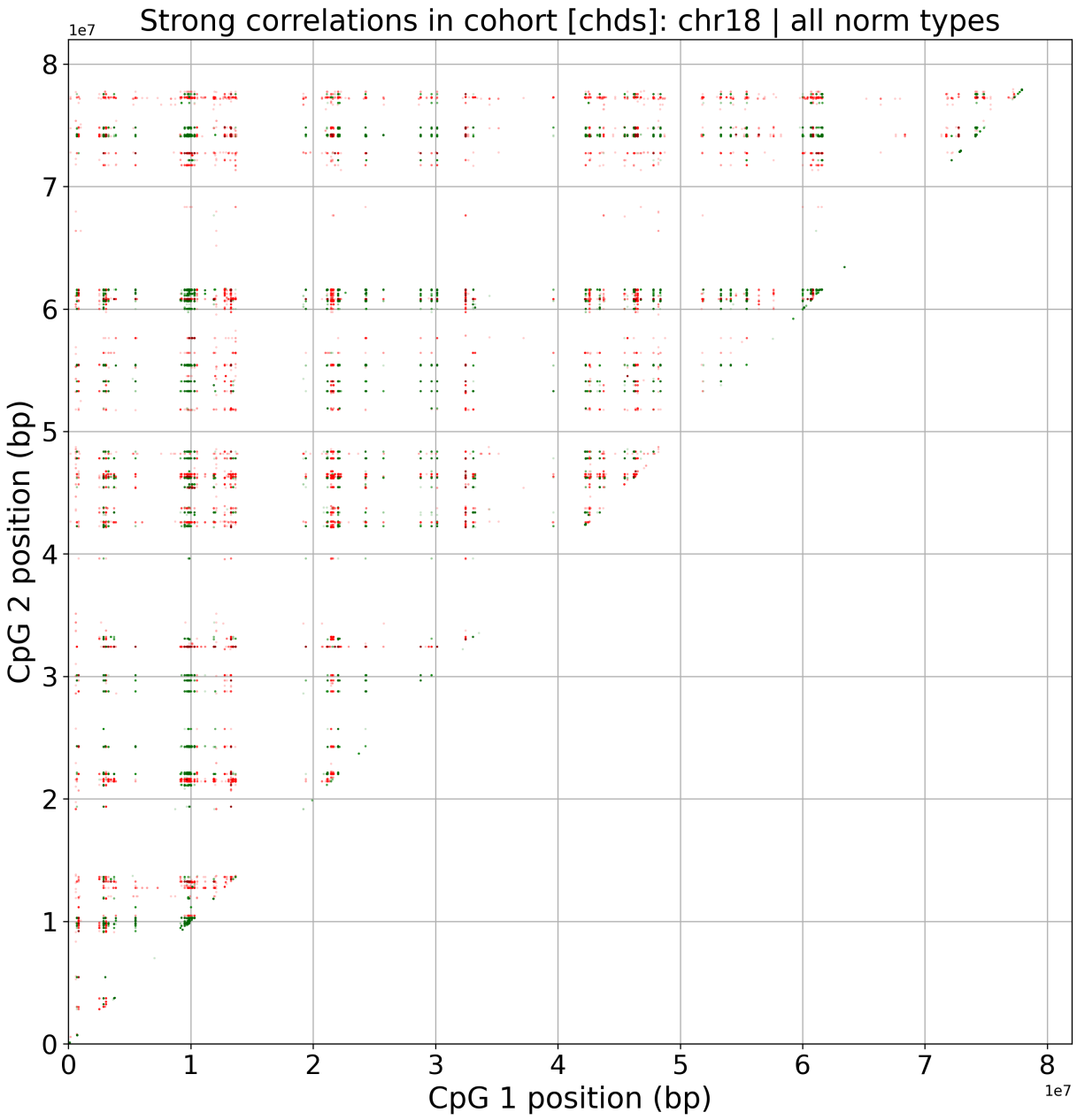

Figure 6.8: Location of strong correlations in CpG methylation intensity (beta) for chromosome 18, with intensity based on score for the metaN method. Source data: CHDS dataset

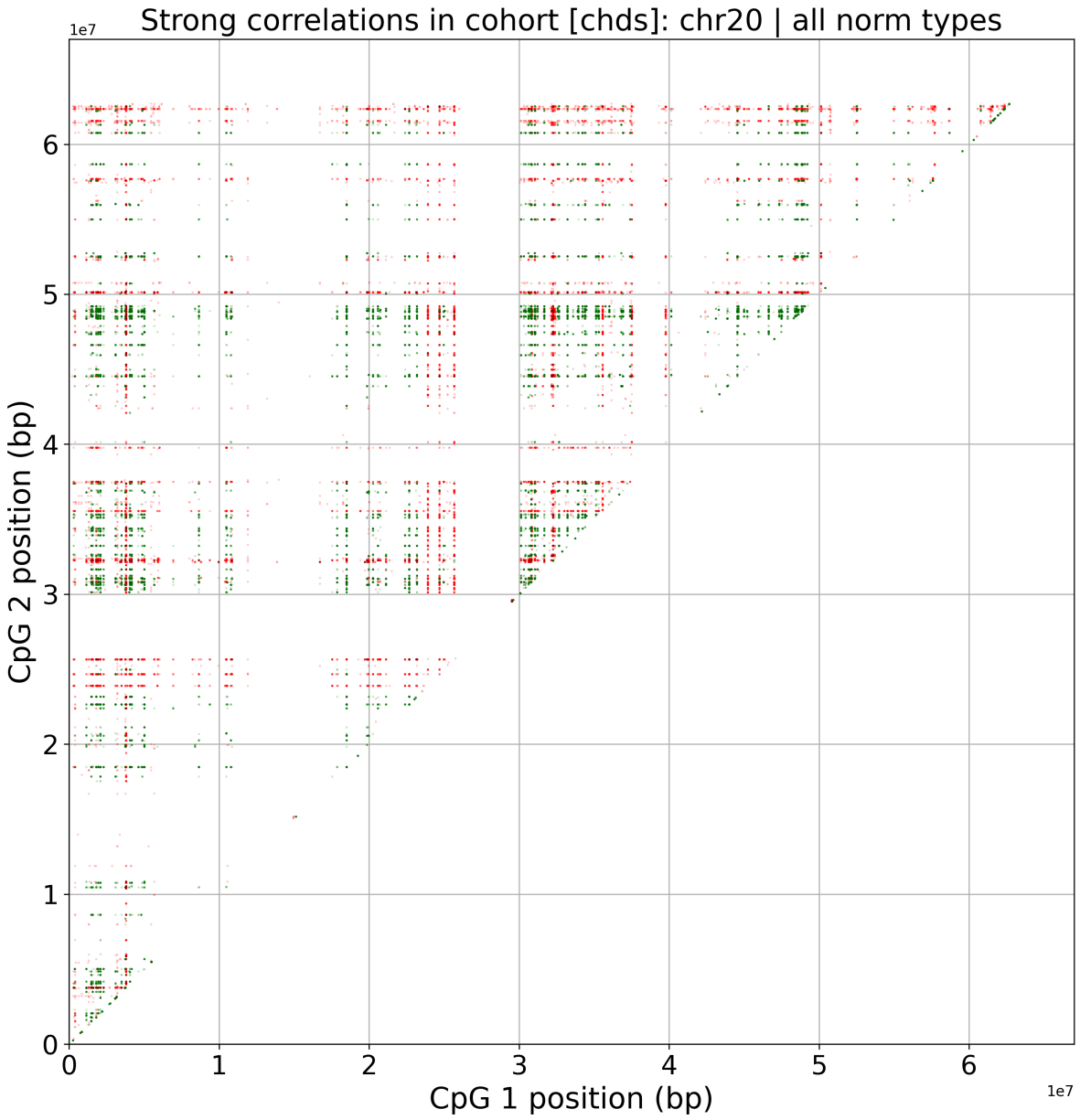

Figure 6.9: Location of strong correlations in CpG methylation intensity (beta) for chromosome 20, with intensity based on score for the metaN method. Source data: CHDS dataset

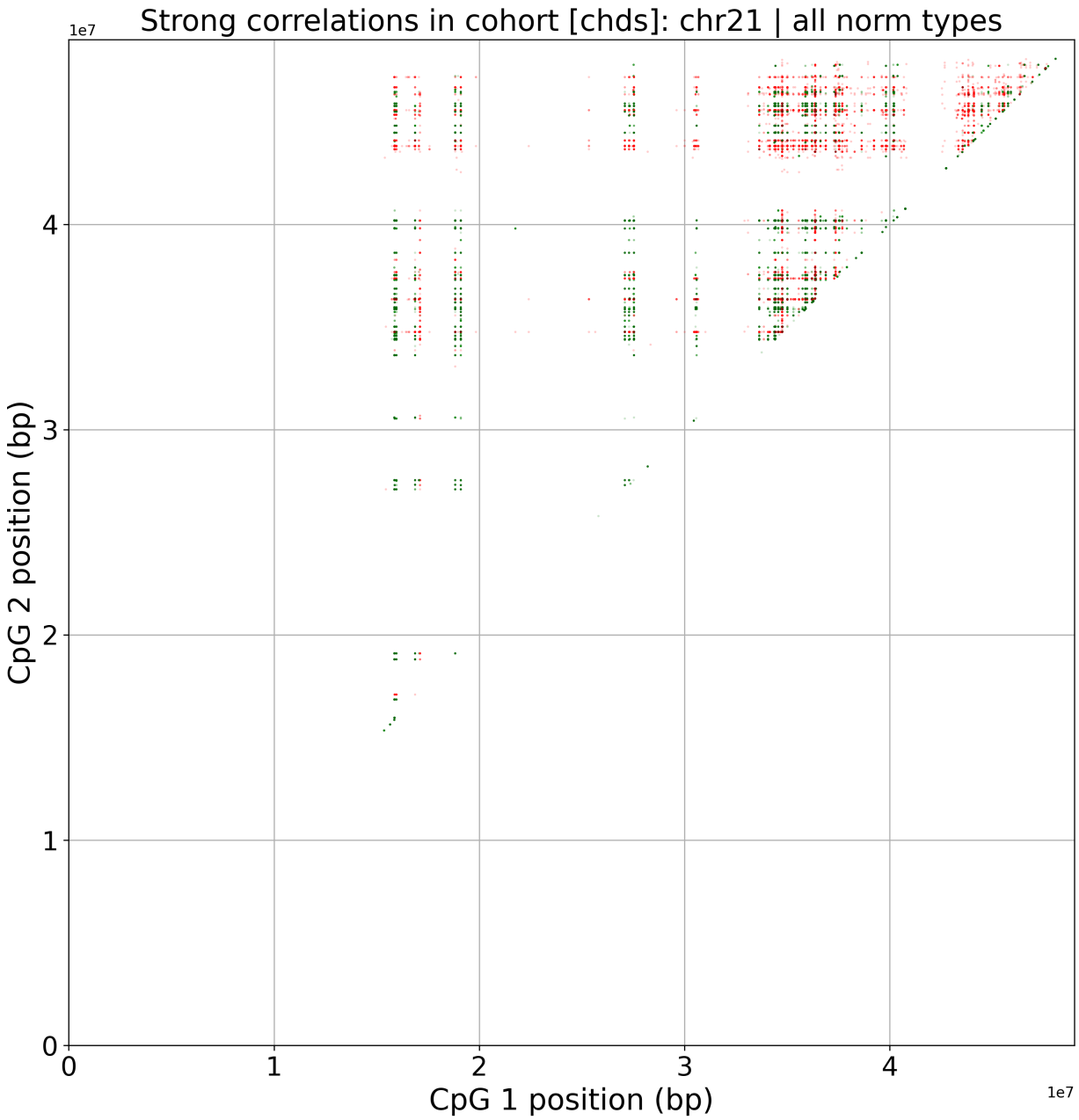

Figure 6.10: Location of strong correlations in CpG methylation intensity (beta) for chromosome 21, with intensity based on score for the metaN method. Source data: CHDS dataset

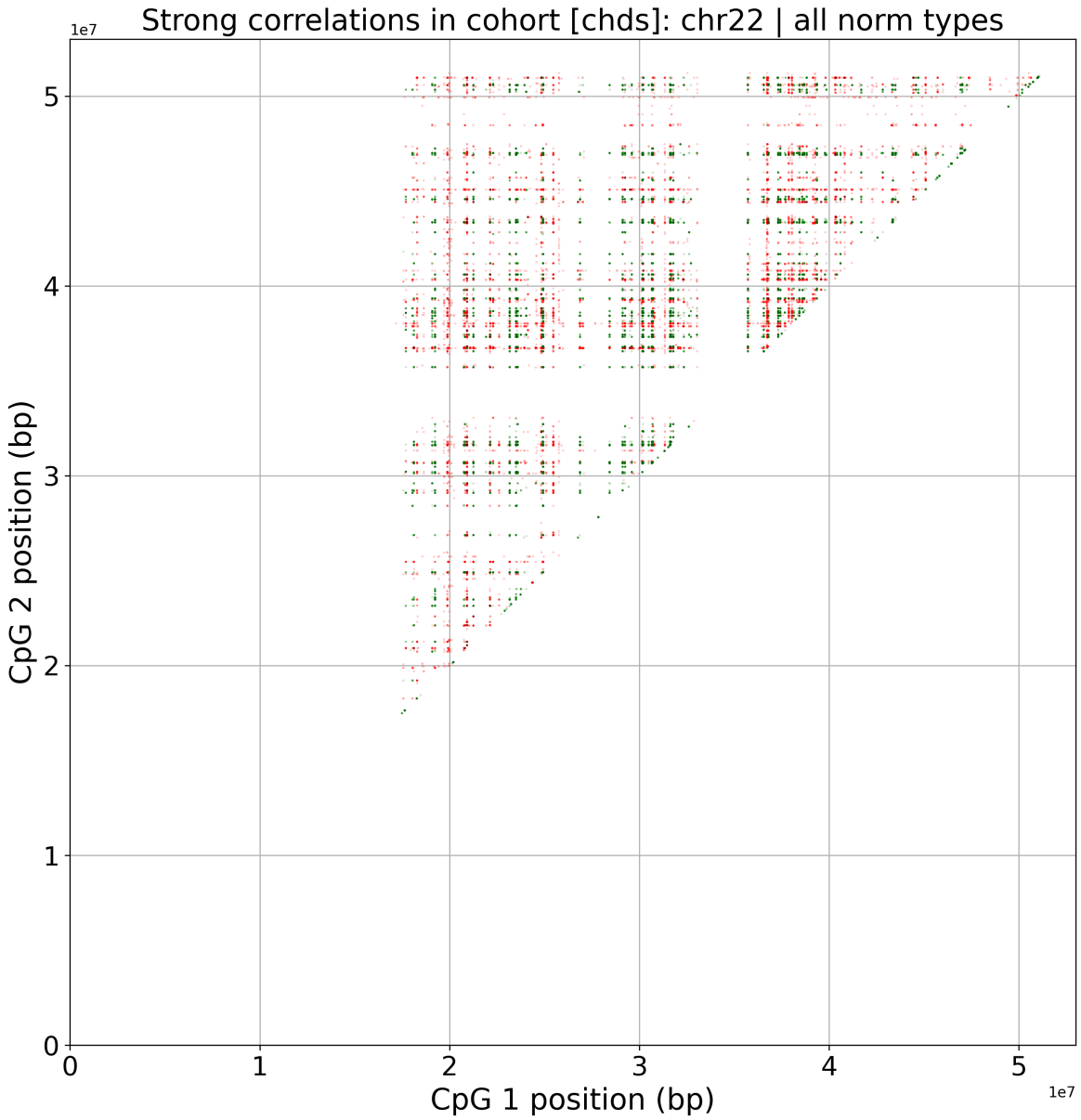

Figure 6.11: Location of strong correlations in CpG methylation intensity (beta) for chromosome 22, with intensity based on score for the metaN method. Source data: CHDS dataset

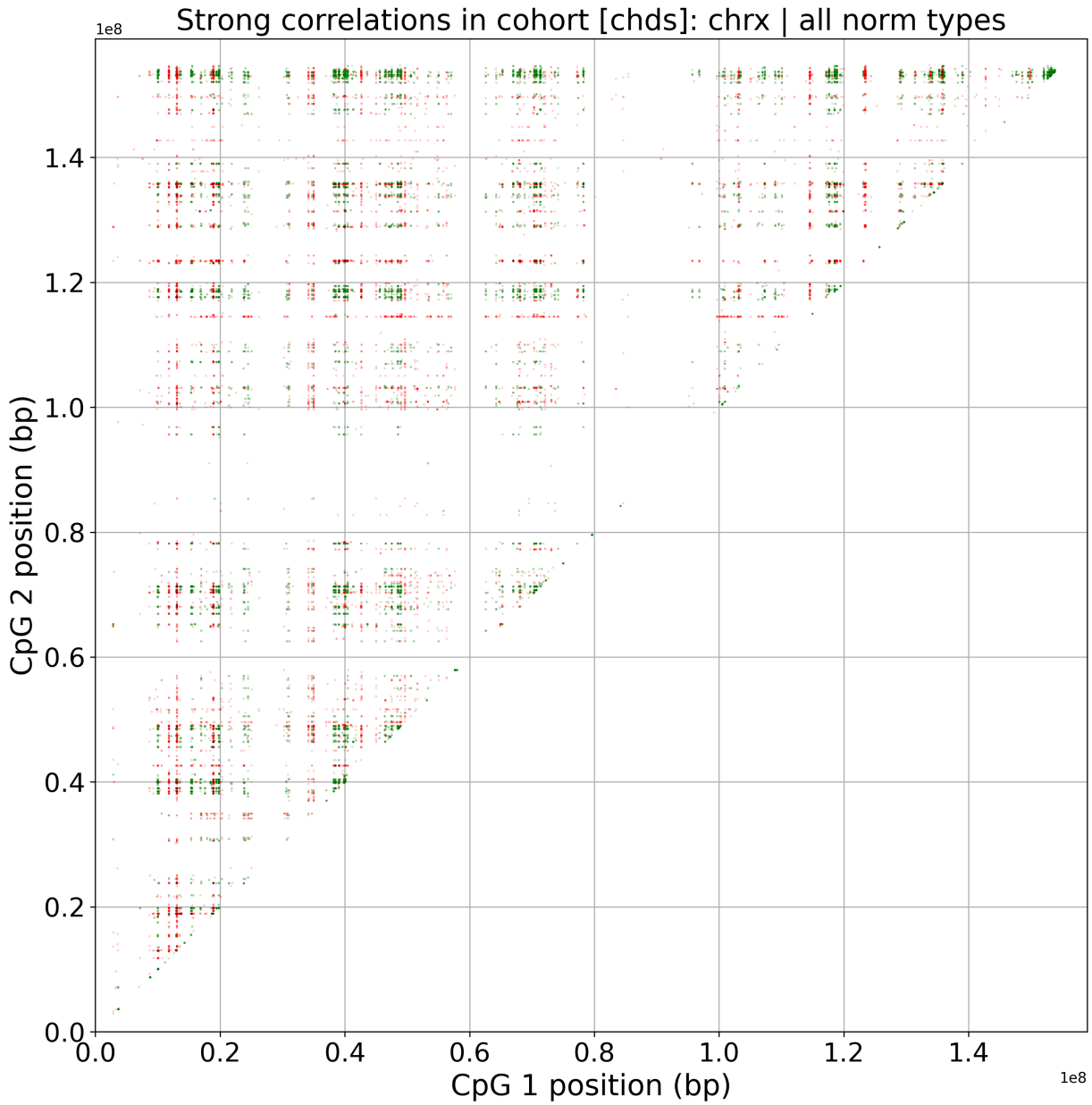

Figure 6.12: Location of strong correlations in CpG methylation intensity (beta) for chromosome X, with intensity based on score for the metaN method. Source data: CHDS dataset

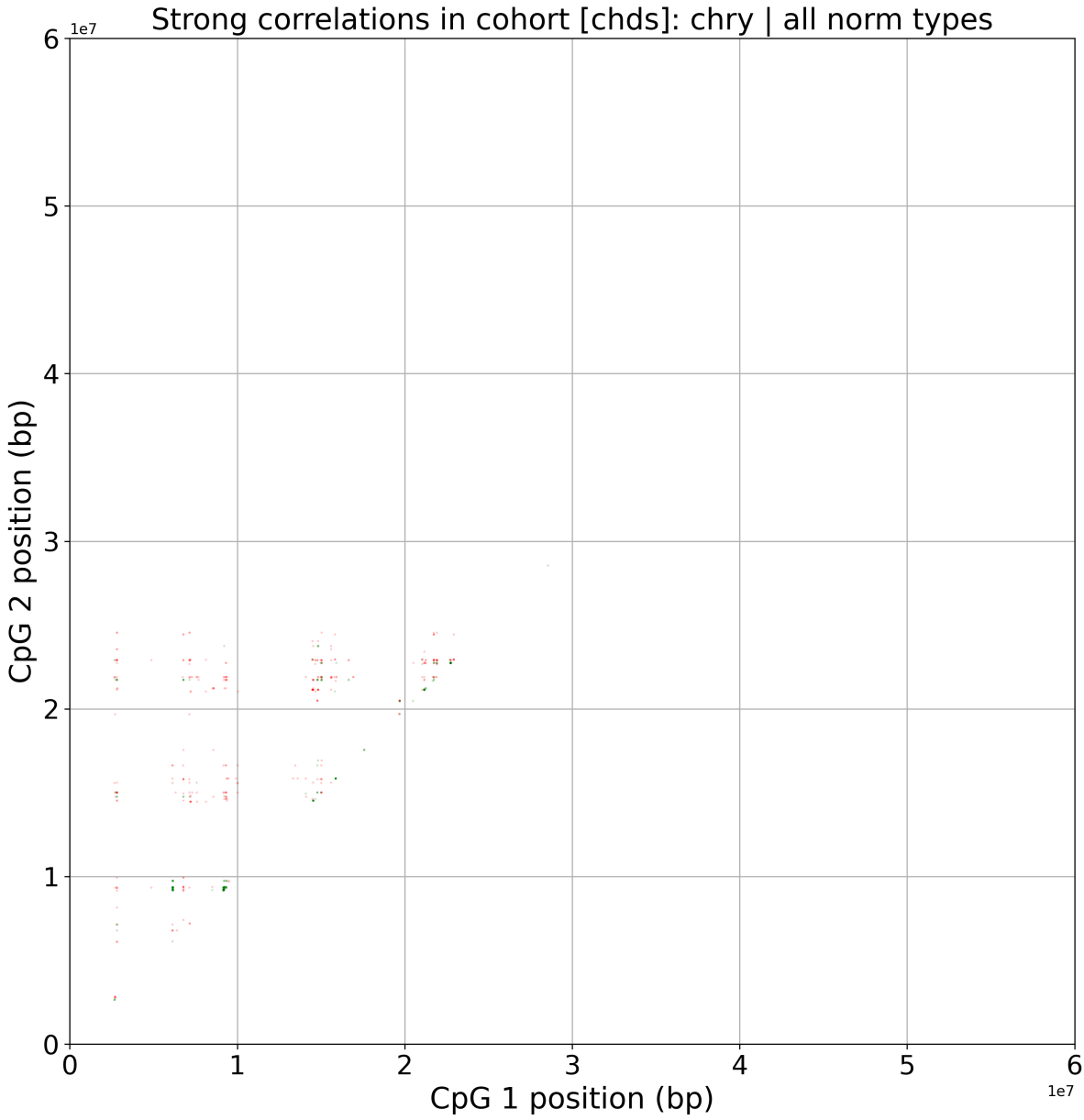

Figure 6.13: Location of strong correlations in CpG methylation intensity (beta) for chromosome Y, with intensity based on score for the metaN method. Source data: CHDS dataset

In all tested chromosomes, we see a distinct banding pattern across both axes. This is accompanied by notable clusters in the data where we see an abundance of strong correlations (either positive or negative) in a localised area. We also see a positively-correlating diagonal for all chromosomes. This indicates a tendency for strong positive correlations to occur in 'data-adjacent' CpGs, though we must keep in mind that there may be unprobed CpGs between CpG sites in our data as the EPIC array only observes 3% of the CpG sites in the human genome.

6.5 Study: Correlation trends within CpG islands

CpG islands, discussed in section 1.2, are loosely-defined regions of the genome that contain a high frequency of CpG sites. This results in increased proximity between sites; though spatial effects throughout the genome are assessed more generally in other studies in this thesis, this study focuses specifically on regions that have been annotated as CpG islands. Results from section 6.4 show that CpG islands may contain strong positive correlations, and these are visible as a diagonal in the location plot of CpG pairs for each chromosome. In this study, we quantify these findings using statistical methods.

6.5.1 Methods

Methods from section 2.2 are used to obtain correlation matrices for a selection of chromosomes. The computational intensity of this procedure is a limiting factor (discussed in section 2.3.1) so only the following chromosomes are investigated: 4, 8, 9, 13, 14, 15, 16, 18, 20, 21, 22, X and Y.

CpG sites are associated with islands based on the Illumina manifest, which contains information on CpG islands defined by UCSC. We extract a list of all annotated islands and their associated CpG sites for each chromosome. CpG sites from islands that have more than one associated CpG site are used to produce two sets:

- A set of all possible pairs within each island, with $\frac{n!}{2!(n-2)!}$ pairs possible for n CpG sites within an island
- A set of all data-adjacent pairs within each island, i.e. the CpG sites are sorted by distance and pairs taken from adjacent elements in this sorted set, with $n-1$ pairs possible for n CpG sites within an island

Both of these sets are divided into positive and negative subsets then subjected to one-way ANOVA against all correlations within a chromosome to test for the statistical significance of their differences. Python's *scipy* module is used for ANOVA functionality.

6.5.2 Results

	Whole chromosome	**All CpG island pairs**	**Adjacent CpG island pairs**
Number of positive correlations	403073784	74658	10554
Number of negative correlations	272961051	37451	2595
Mean positive correlation	0.146	0.17	0.243
Mean negative correlation	-0.103	-0.0992	-0.0838
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	2670.0	5810.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	67.5	116.0
	Negatives ANOVA p-value	2.1e-16	3.7e-27

Table 6.35: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 4, and all correlations within chromosome 4 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 1010 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	449077532	77084	10727
Number of negative correlations	290181394	39323	2569
Mean positive correlation	0.139	0.17	0.245
Mean negative correlation	-0.0986	-0.0948	-0.0851
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	4820.0	7960.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	72.9	60.8
	Negatives ANOVA p-value	1.35e-17	6.29e-15

Table 6.36: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 8, and all correlations within chromosome 8 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 1001 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	210811274	45242	8246
Number of negative correlations	131531587	24096	2151
Mean positive correlation	0.133	0.162	0.227
Mean negative correlation	-0.0953	-0.0941	-0.0827
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	2900.0	5510.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	5.21	47.5
	Negatives ANOVA p-value	0.0224	5.58e-12

Table 6.37: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 9, and all correlations within chromosome 9 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 1145 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	136109854	45383	6114
Number of negative correlations	85220426	19952	1389
Mean positive correlation	0.142	0.172	0.246
Mean negative correlation	-0.0997	-0.0933	-0.0866
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	2540.0	4180.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	103.0	29.2
	Negatives ANOVA p-value	3.85e-24	6.57e-08

Table 6.38: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 13, and all correlations within chromosome 13 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 573 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	266185214	71333	9113
Number of negative correlations	170401261	35364	2110
Mean positive correlation	0.135	0.165	0.24
Mean negative correlation	-0.0992	-0.0939	-0.0826
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	4630.0	7240.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	124.0	71.8
	Negatives ANOVA p-value	8.15e-29	2.41e-17

Table 6.39: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 14, and all correlations within chromosome 14 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 779 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	254297067	64196	8491
Number of negative correlations	158711103	33240	2026
Mean positive correlation	0.135	0.164	0.239
Mean negative correlation	-0.0971	-0.095	-0.0839
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	3970.0	6830.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	19.5	47.0
	Negatives ANOVA p-value	9.95e-06	7.02e-12

Table 6.40: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 15, and all correlations within chromosome 15 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 759 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	449528374	106536	14661
Number of negative correlations	270136517	50734	3256
Mean positive correlation	0.125	0.165	0.24
Mean negative correlation	-0.0928	-0.0922	-0.0804
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	15600.0	17900.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	2.15	69.9
	Negatives ANOVA p-value	0.143	6.14e-17

Table 6.41: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 16, and all correlations within chromosome 16 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 1433 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	68742921	31732	4557
Number of negative correlations	42239730	14625	1060
Mean positive correlation	0.143	0.173	0.252
Mean negative correlation	-0.0995	-0.0921	-0.082
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	1860.0	3480.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	101.0	41.5
	Negatives ANOVA p-value	1.13e-23	1.2e-10

Table 6.42: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 18, and all correlations within chromosome 18 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 504 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	165890096	59054	8096
Number of negative correlations	97679224	27088	1723
Mean positive correlation	0.131	0.187	0.262
Mean negative correlation	-0.0936	-0.0921	-0.081
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	15200.0	11200.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	9.05	38.9
	Negatives ANOVA p-value	0.00263	4.49e-10

Table 6.43: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 20, and all correlations within chromosome 20 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 800 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	33906857	22451	3368
Number of negative correlations	19132993	9861	691
Mean positive correlation	0.137	0.177	0.252
Mean negative correlation	-0.0951	-0.0913	-0.0795
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	2640.0	3320.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	18.5	21.9
	Negatives ANOVA p-value	1.71e-05	2.89e-06

Table 6.44: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 21, and all correlations within chromosome 21 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 352 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	108741296	46751	6899
Number of negative correlations	59922865	21570	1488
Mean positive correlation	0.13	0.177	0.252
Mean negative correlation	-0.0921	-0.0919	-0.079
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	9010.0	9010.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	0.192	35.6
	Negatives ANOVA p-value	0.661	2.48e-09

Table 6.45: Statistical comparison of correlations between CpG sites on CpG islands on chromosome 22, and all correlations within chromosome 22 in general, fromcorrelation matrices derived from noob-normalised beta values. Number of islands with more than one CpG site: 710 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	97951388	47234	7468
Number of negative correlations	84253117	18763	1187
Mean positive correlation	0.361	0.508	0.544
Mean negative correlation	-0.322	-0.334	-0.258
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	21700.0	5280.0
	Positives ANOVA p-value	0.0	0.0
	Negatives ANOVA F-score	84.7	142.0
	Negatives ANOVA p-value	3.48e-20	9.58e-33

Table 6.46: Statistical comparison of correlations between CpG sites on CpG islands on chromosome x, and all correlations within chromosome x in general, fromcorrelation matrices derived from swan-normalised beta values. Number of islands with more than one CpG site: 754 (cohort: chds)

	Whole chromosome	All CpG island pairs	Adjacent CpG island pairs
Number of positive correlations	83330	852	245
Number of negative correlations	60586	369	36
Mean positive correlation	0.442	0.55	0.577
Mean negative correlation	-0.371	-0.386	-0.3
		All islands vs. chromosome	**Adjacent vs. chromosome**
	Positives ANOVA F-score	277.0	126.0
	Positives ANOVA p-value	4.63e-62	4.02e-29
	Negatives ANOVA F-score	2.66	6.23
	Negatives ANOVA p-value	0.103	0.0126

Table 6.47: Statistical comparison of correlations between CpG sites on CpG islands on chromosome y, and all correlations within chromosome y in general, fromcorrelation matrices derived from swan-normalised beta values. Number of islands with more than one CpG site: 59 (cohort: chds)

Our results showed that all of our tested chromosomes have a higher average positive correlation within CpG islands than the baseline level of the entire chromosome. Similarly, data-adjacent CpG pairs have a higher

average positive correlation than the already-elevated average of the CpG island subset. In both cases, results are generally statistically significant as per ANOVA ($p << 0.001$), but this trend of statistical significance is only consistent for strong positive correlations within the autosomes. Mean negative correlation tends to become more positive in most cases with the notable exception of the sex chromosomes - for chromosomes X and Y, the mean negative correlation across all CpG island pairs was more negative than the chromosomal average, but more positive for adjacent CpG island pairs.

Chapter 7

Correlations within genes and pathways

7.1 Premise

CpG sites may strongly correlate (either positively or negatively) in methylation intensity with multiple other CpG sites. A key hypothesis that we must test is the methylation intensity of CpG sites within genes and biological pathways will tend to correlate more-strongly than those not on these similar functional groups.

In this chapter, we undertake two studies:

- An analysis of the statistical distribution of correlations within the same genes, versus that of correlations across the chromosome in general
- An analysis of the difference in distribution of correlations within the same pathways, versus that of correlations across the chromosome in general

The results in this chapter are best considered in conjunction with the results from chapter 6. Both chapters are discussed in a combined general discussion in chapter 8.

7.2 Study: Correlation trends within genes

In this study, we aim to find evidence of significant correlation in methylation intensity between different CpG sites within the same gene. To achieve, this, we estimate a probability distribution function (in the form of a histogram) of correlation coefficients for correlations within genes, versus that of the overall beta correlation matrix.

7.2.1 Methods

Methods from section 2.2 are used to obtain correlation matrices for a selection of chromosomes. The computational intensity of this procedure is a limiting factor (discussed in section 2.3.1) so only the following chromosomes are investigated: 4, 8, 9, 13, 14, 15, 16, 18, 20, 21, 22, X and Y. As per section 2.2.3, CpG sites are related to genes using the Illumina manifest. Two overlapping histograms are generated using the

Python *matplotlib* module - one for the distribution of correlation coefficients across the entire chromosome, and one for the distribution of correlation coefficients between CpG pairs located on the same annotated gene. Distributions are approximated by scaling the total number of items in each histogram bin by the total number of items in that histogram.

A second set of results is produced using the same methods as above, except we exclude CpG islands (as annotated in the Illumina manifest).

7.2.2 Results

7.2.2.1 Distributions with CpG islands

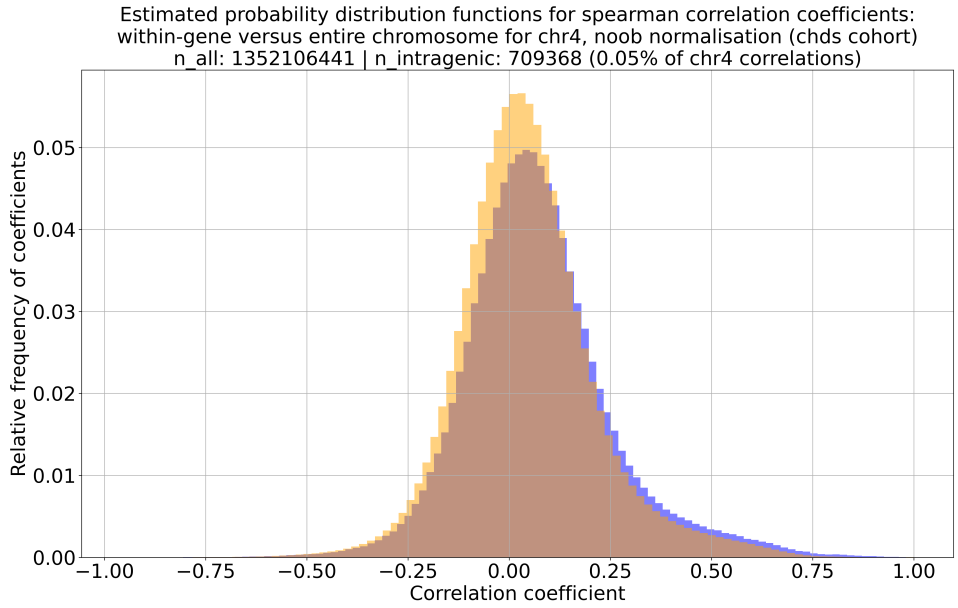

Figure 7.1: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 4. Source data: CHDS dataset

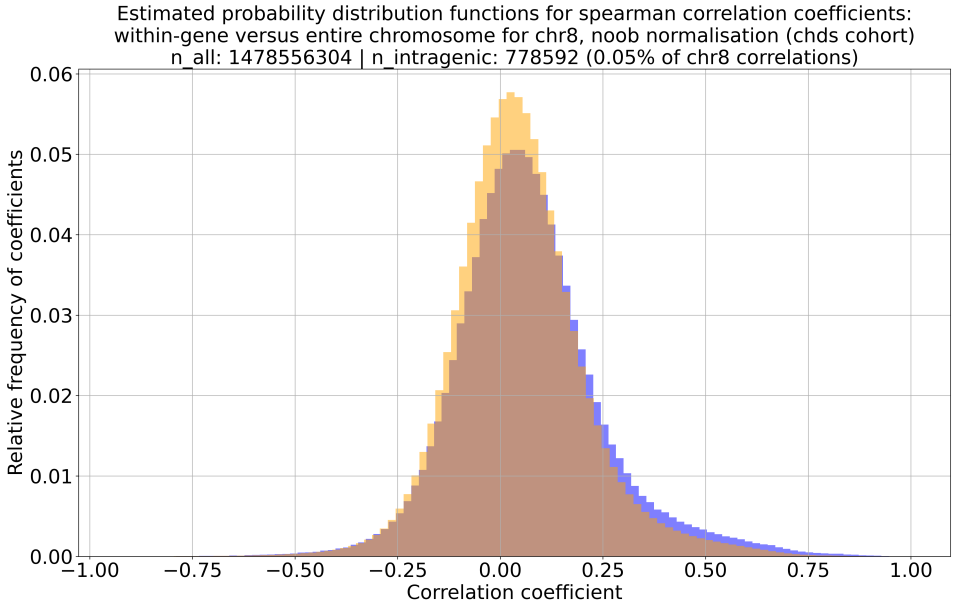

Figure 7.2: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 8. Source data: CHDS dataset

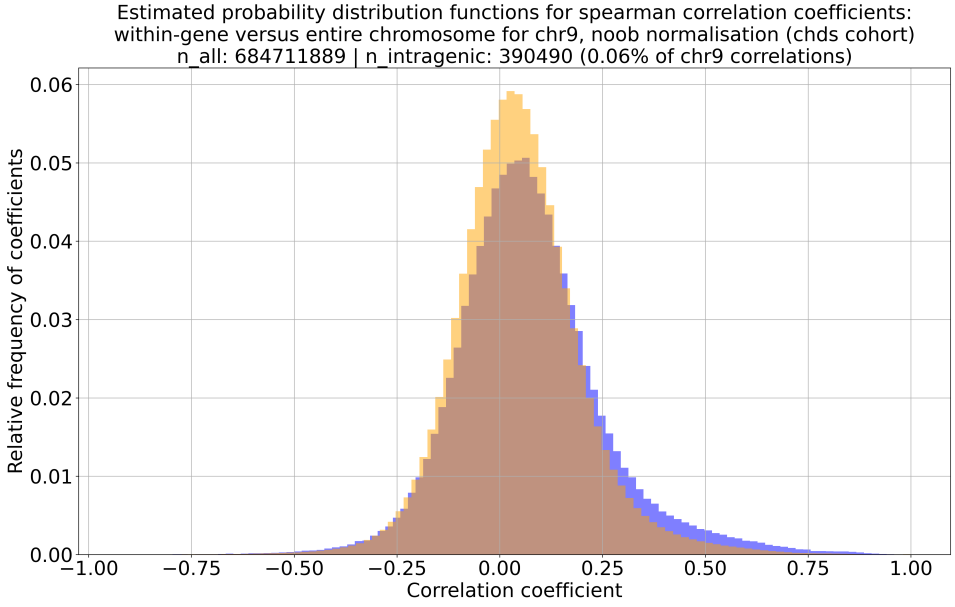

Figure 7.3: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 9. Source data: CHDS dataset

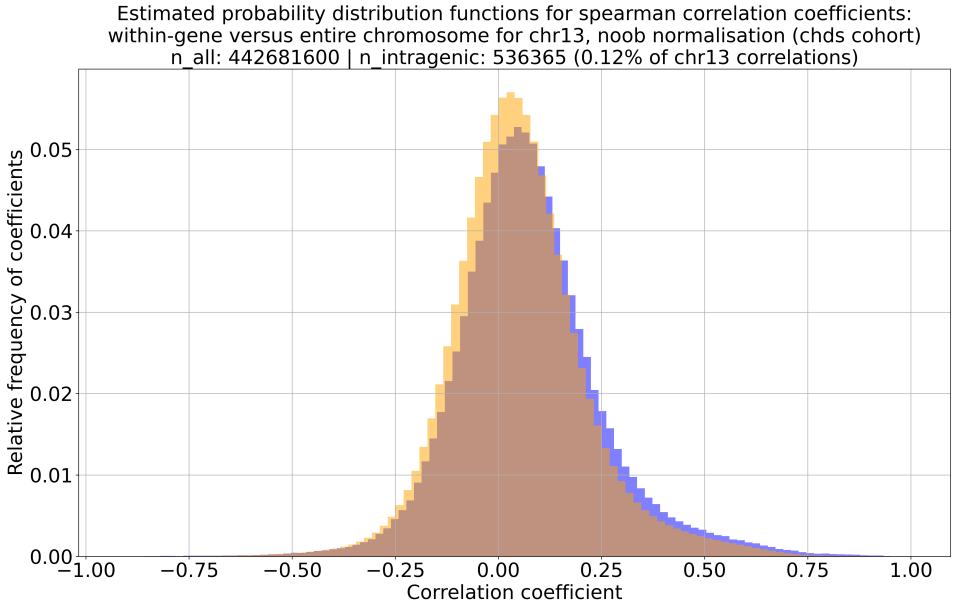

Figure 7.4: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 13. Source data: CHDS dataset

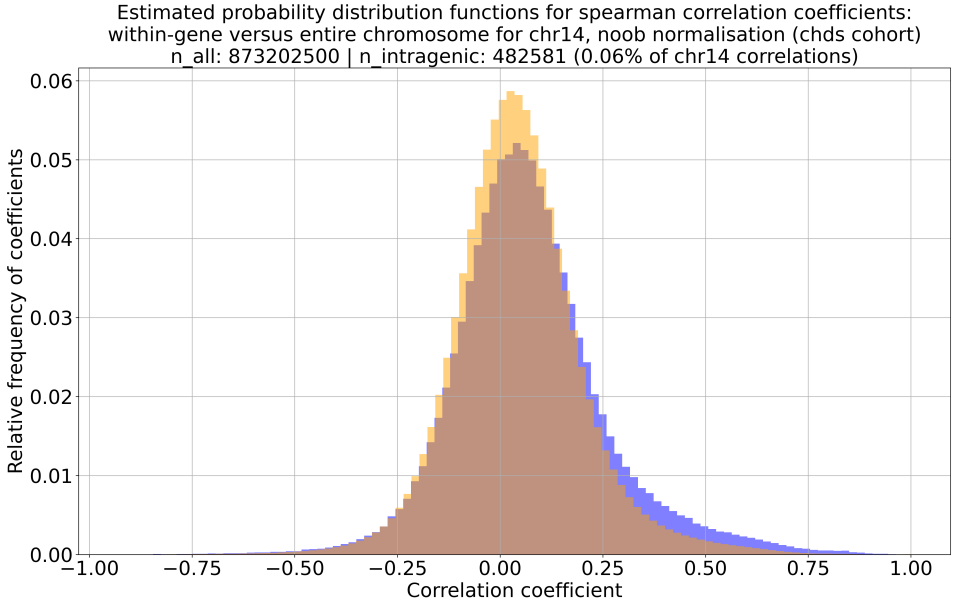

Figure 7.5: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 14. Source data: CHDS dataset

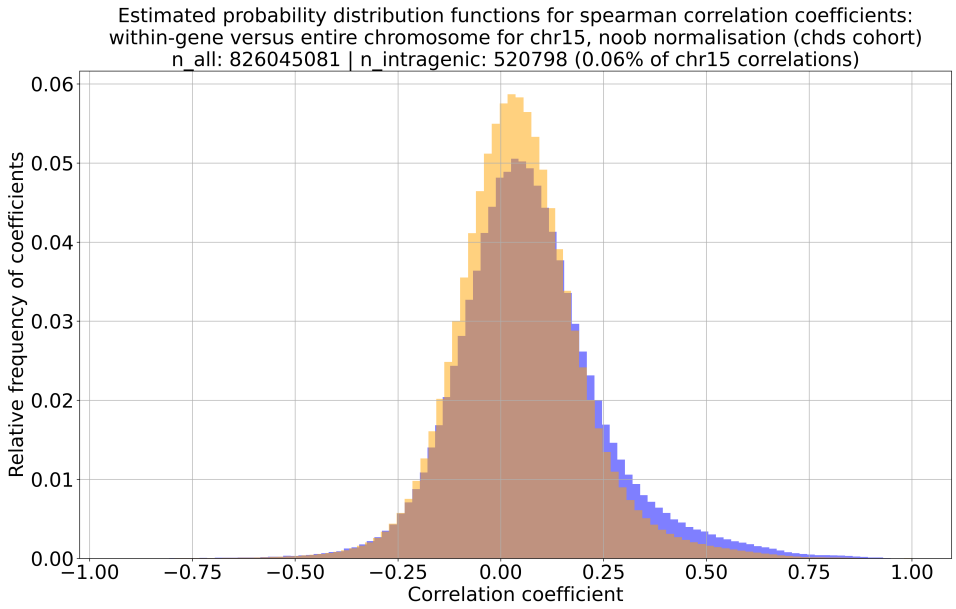

Figure 7.6: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 15. Source data: CHDS dataset

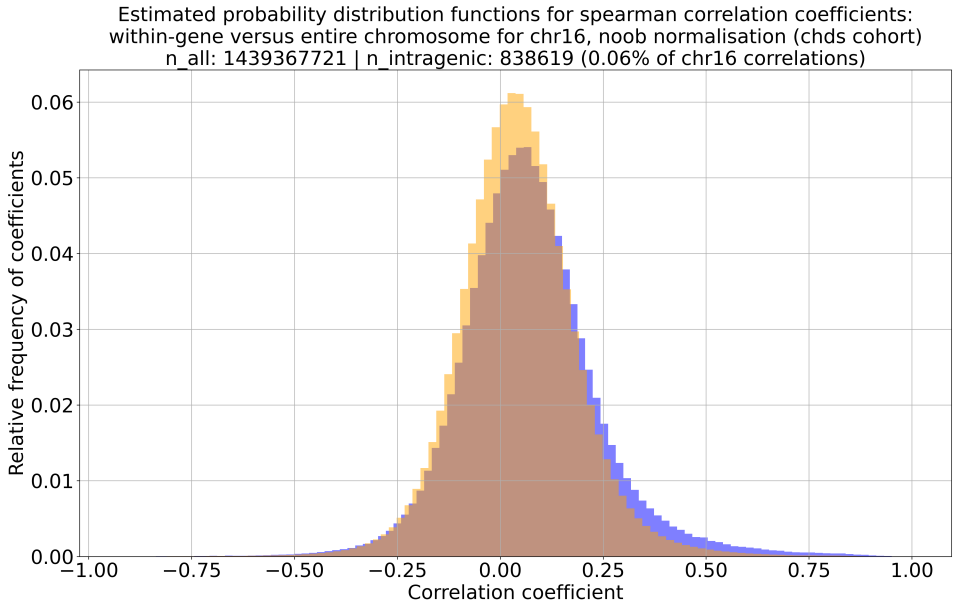

Figure 7.7: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 16. Source data: CHDS dataset

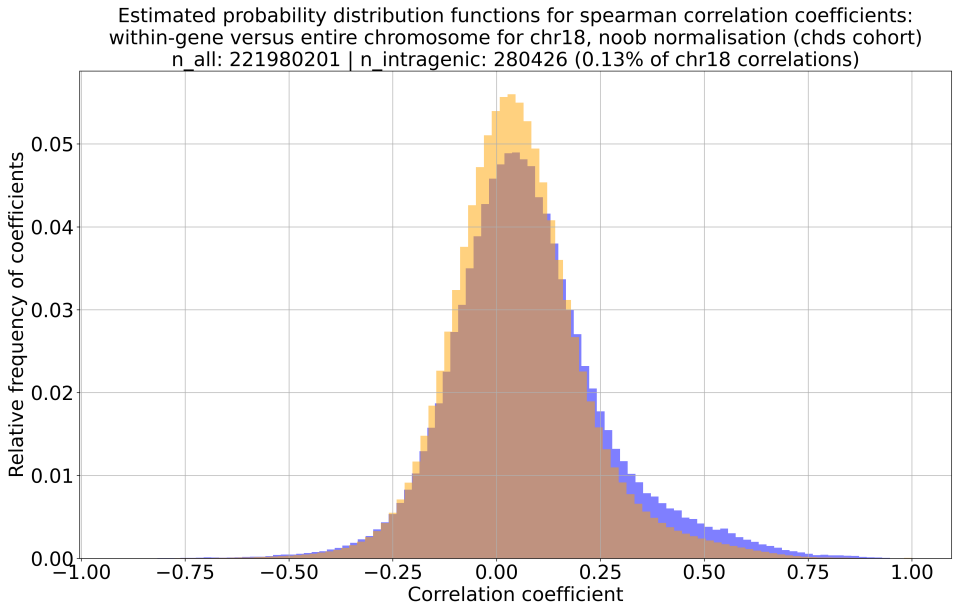

Figure 7.8: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 18. Source data: CHDS dataset

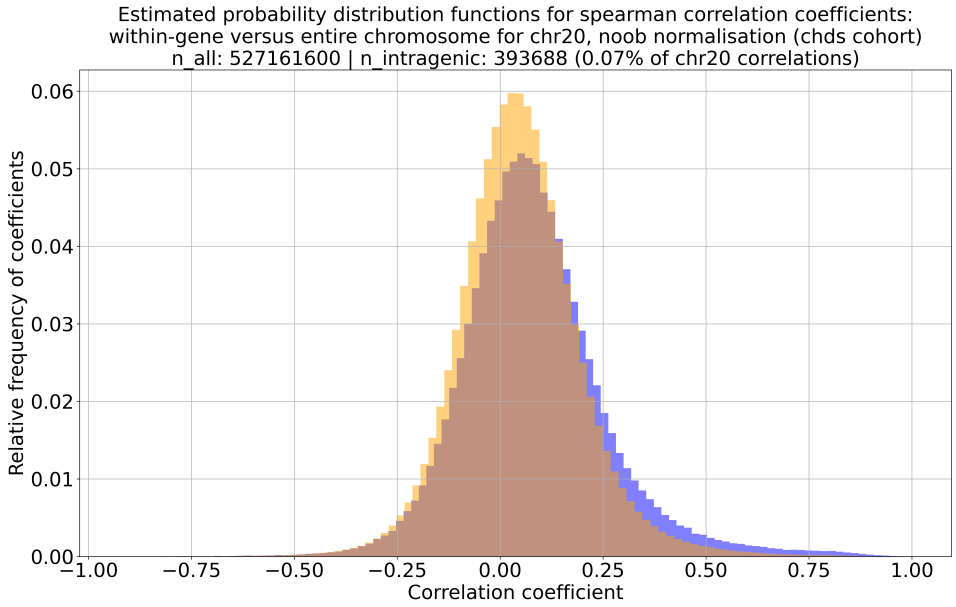

Figure 7.9: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 20. Source data: CHDS dataset

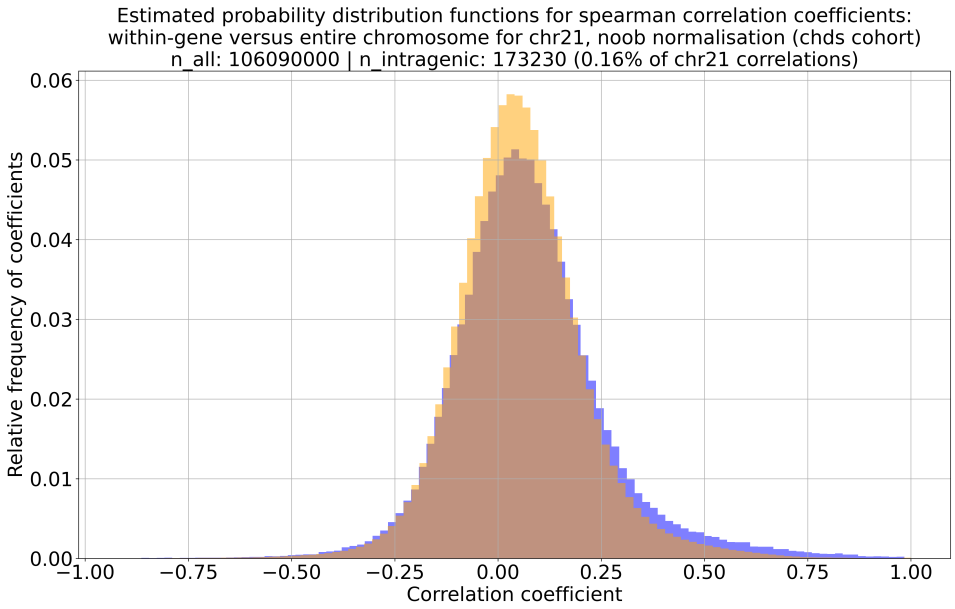

Figure 7.10: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 21. Source data: CHDS dataset

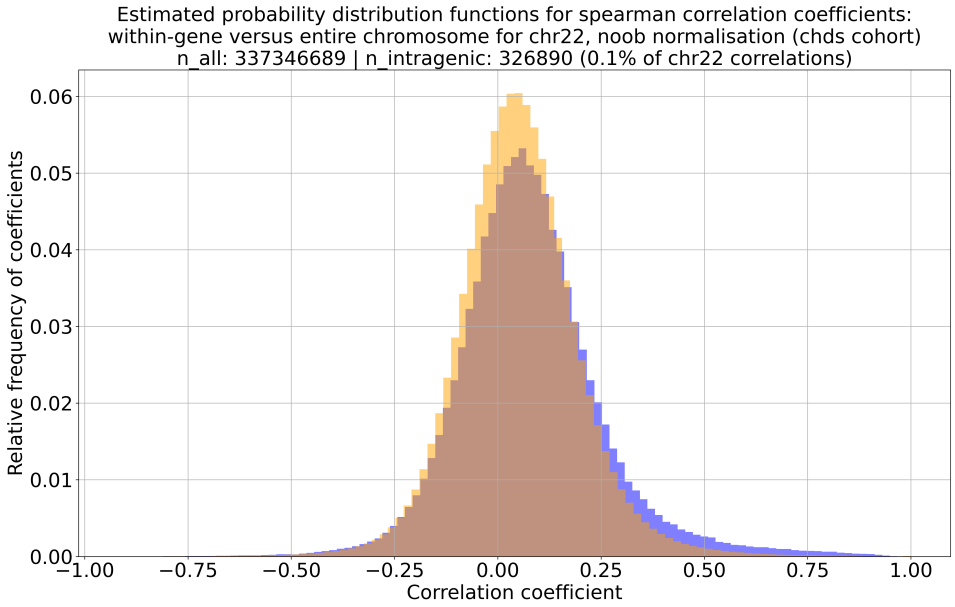

Figure 7.11: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 22. Source data: CHDS dataset

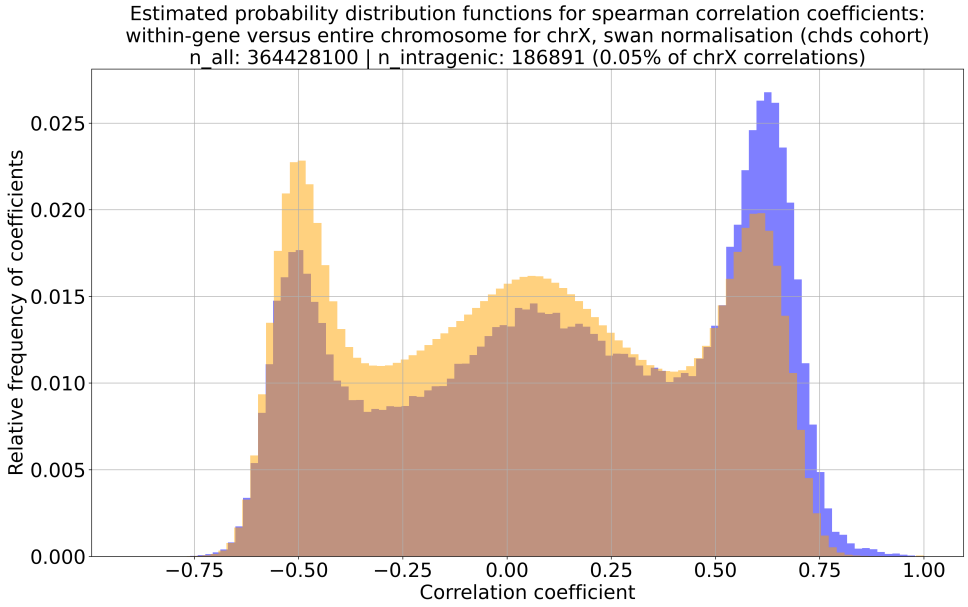

Figure 7.12: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome X. Source data: CHDS dataset

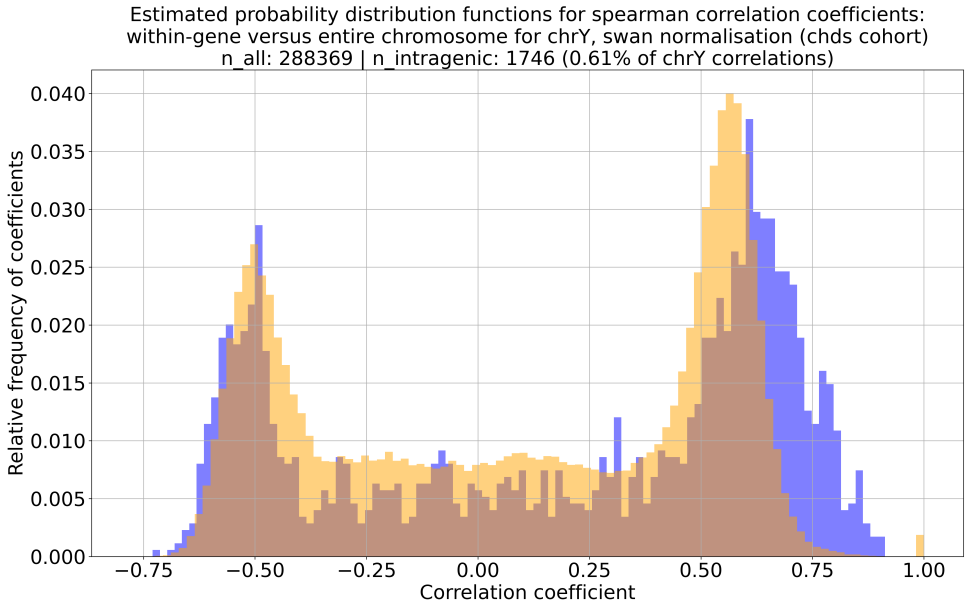

Figure 7.13: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome Y. Source data: CHDS dataset

For all autosomes, we can see a notable section of the estimated distribution function for correlation within the same gene lies to the right of that for the entire chromosome. The right tail is also significantly longer and fatter in these cases, and their peak is also slightly to the right of that of the entire chromosome distribution;

all of this suggests that same-gene correlations tend to have stronger correlations for autosomes. Though hard to see, the left tail of the same-gene distribution is also thicker in some cases, suggesting that there's an increased tendency for stronger negative correlations than the average for these chromosomes.

The same trend is much harder to see in the sex chromosomes (owing to their multimodal, as opposed to unimodal distributions), but as with the autosomes, the distribution of same-gene correlations has a notable component to the right of the whole-chromosome curve.

7.2.2.2 Distributions without CpG islands

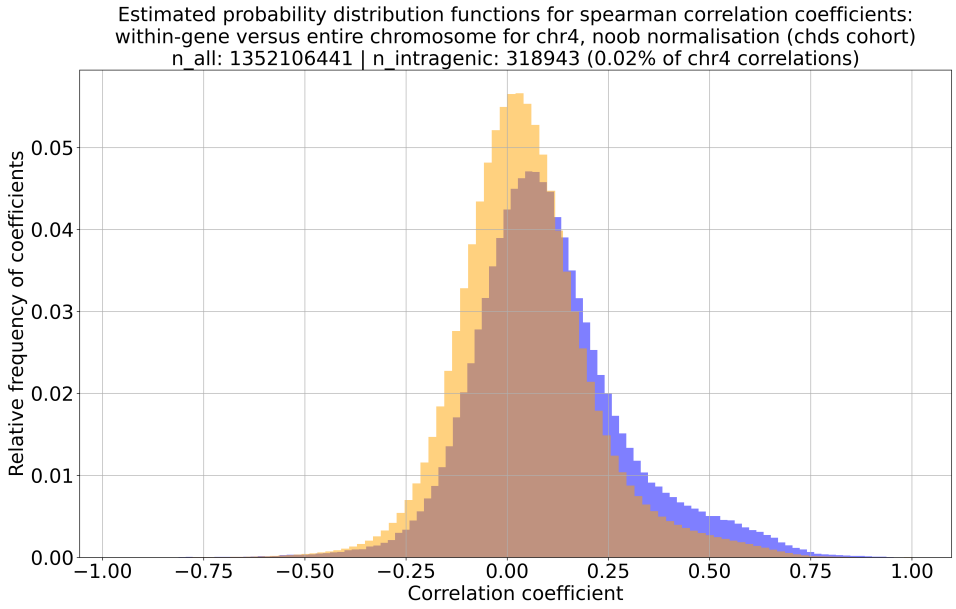

Figure 7.14: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 4, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

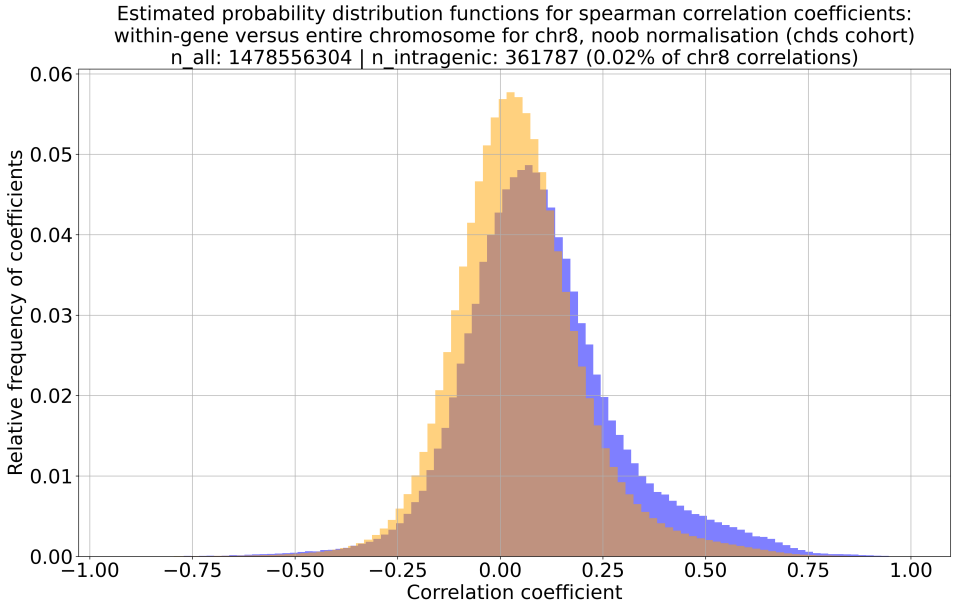

Figure 7.15: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 8, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

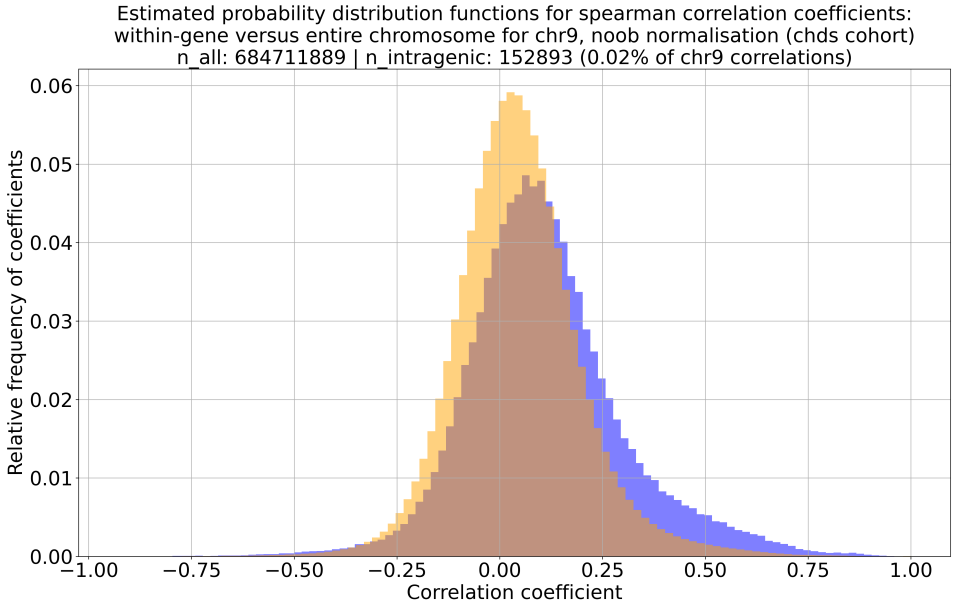

Figure 7.16: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 9, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

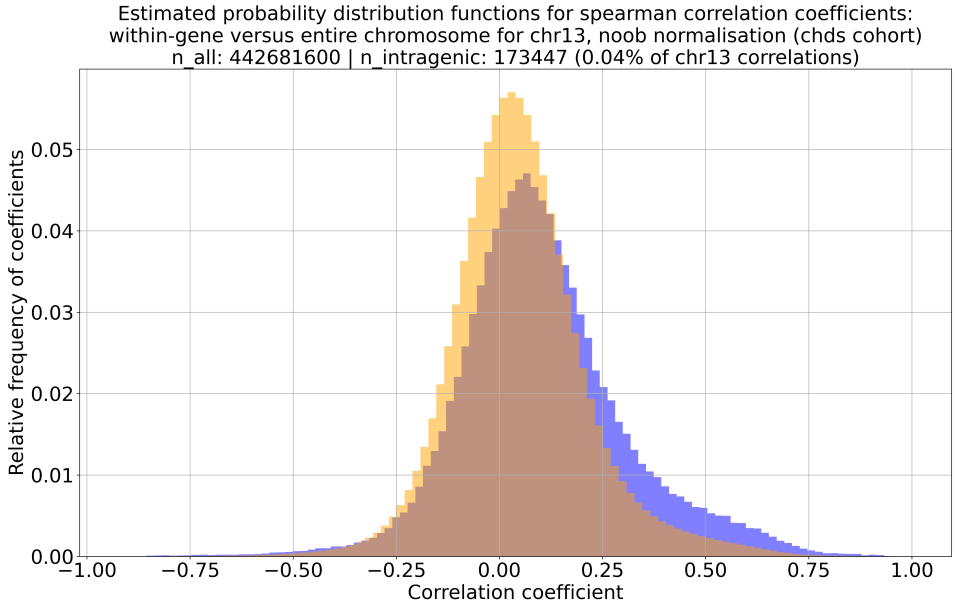

Figure 7.17: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 13, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

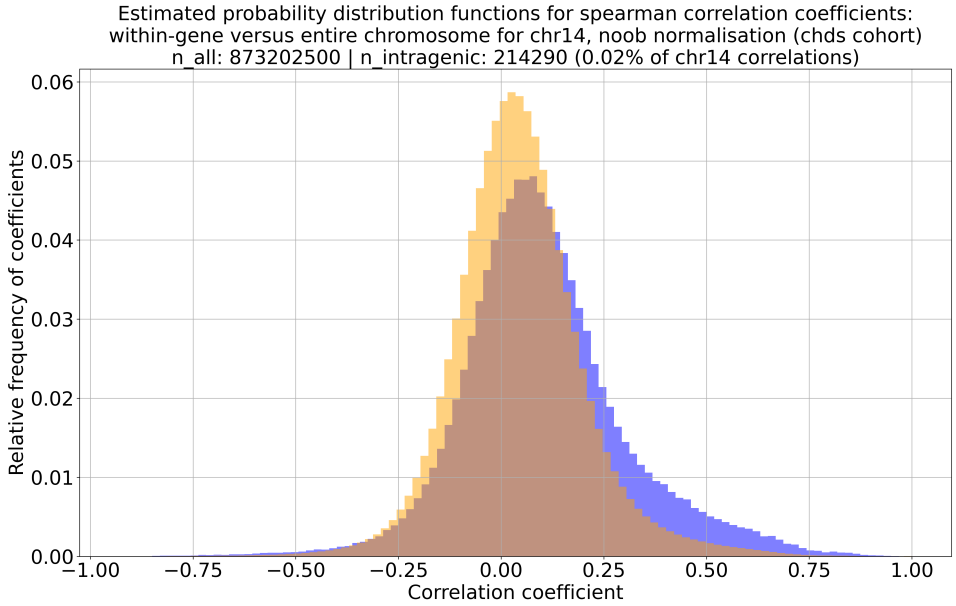

Figure 7.18: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 14, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

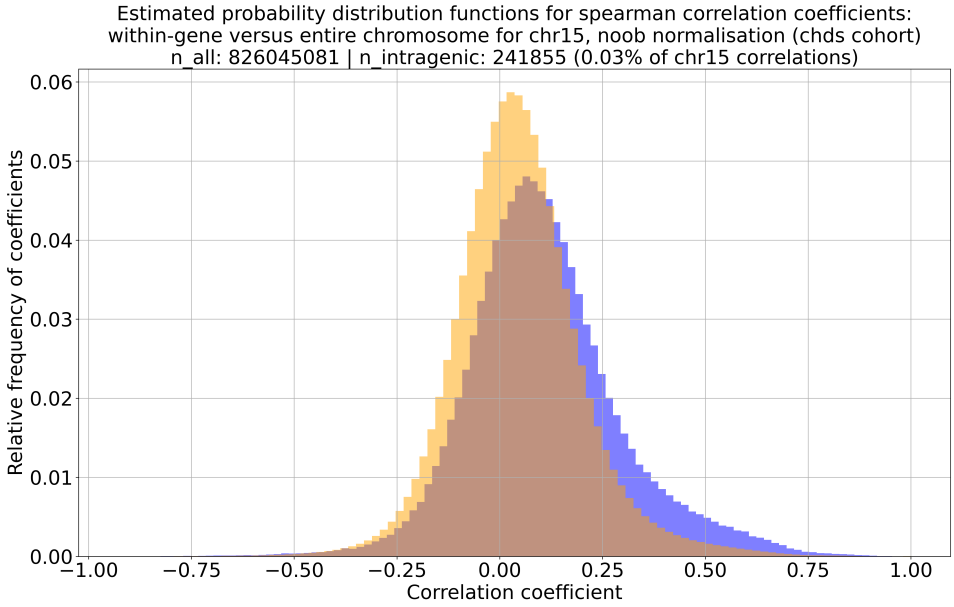

Figure 7.19: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 15, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

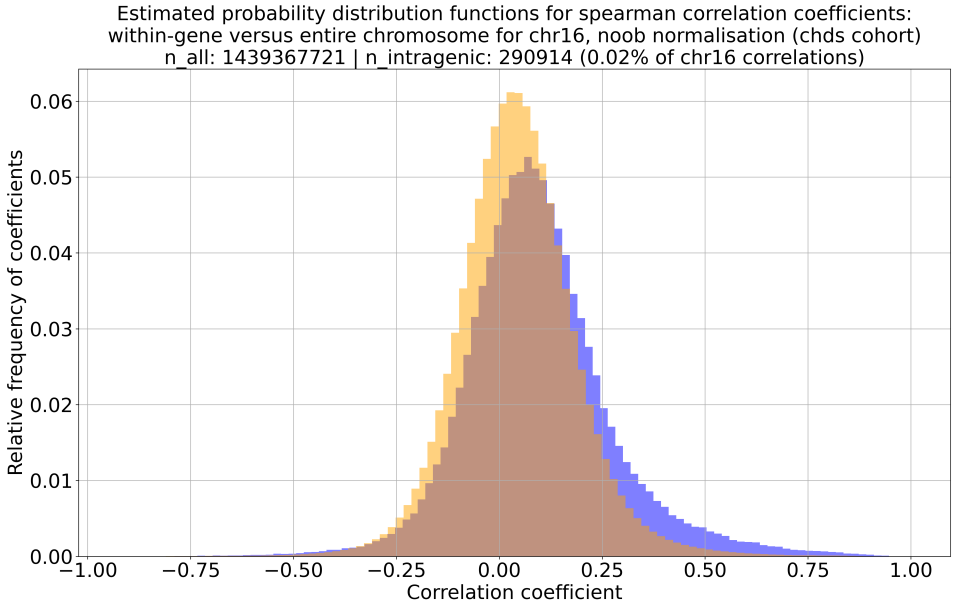

Figure 7.20: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 16, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

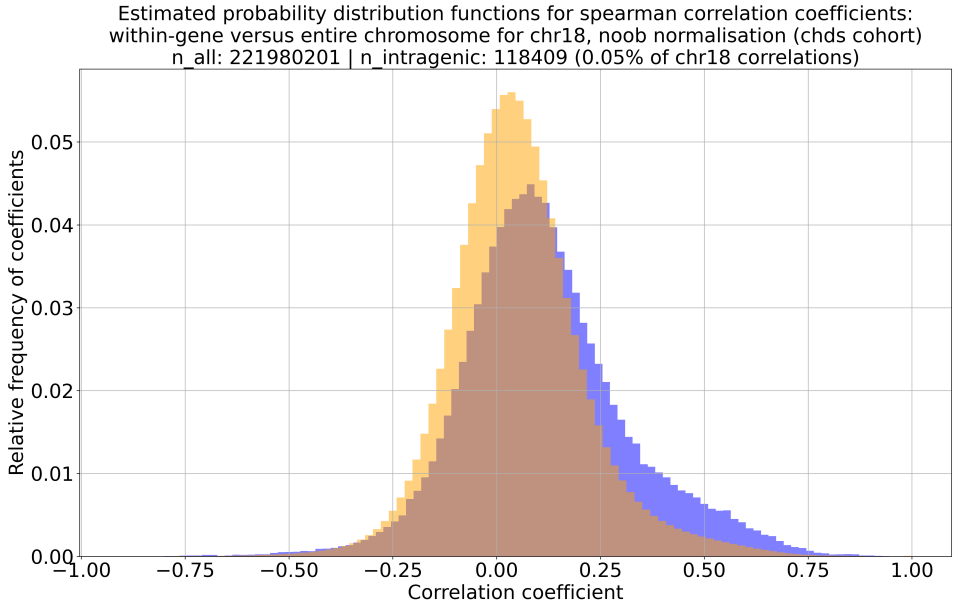

Figure 7.21: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 18, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

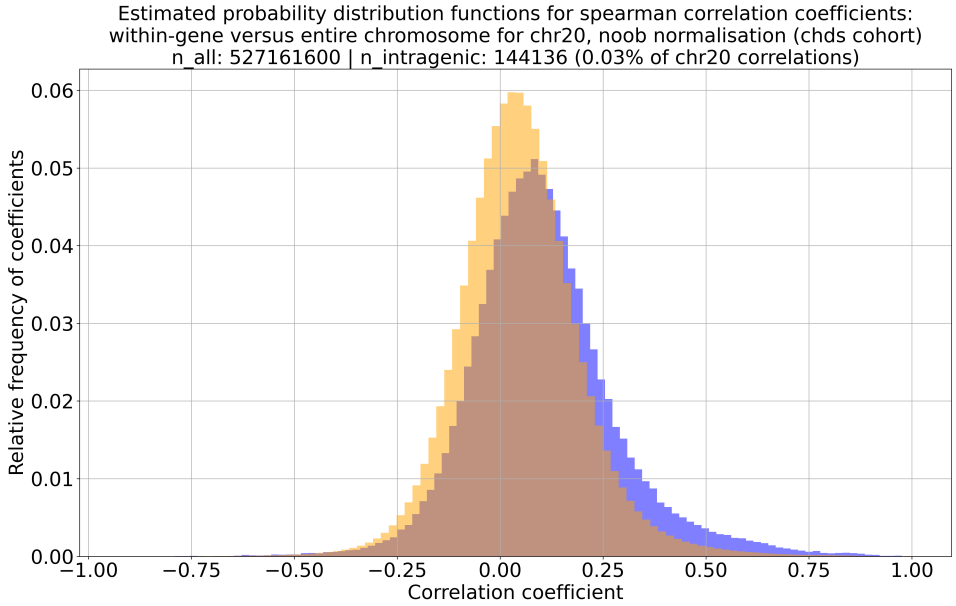

Figure 7.22: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 20, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

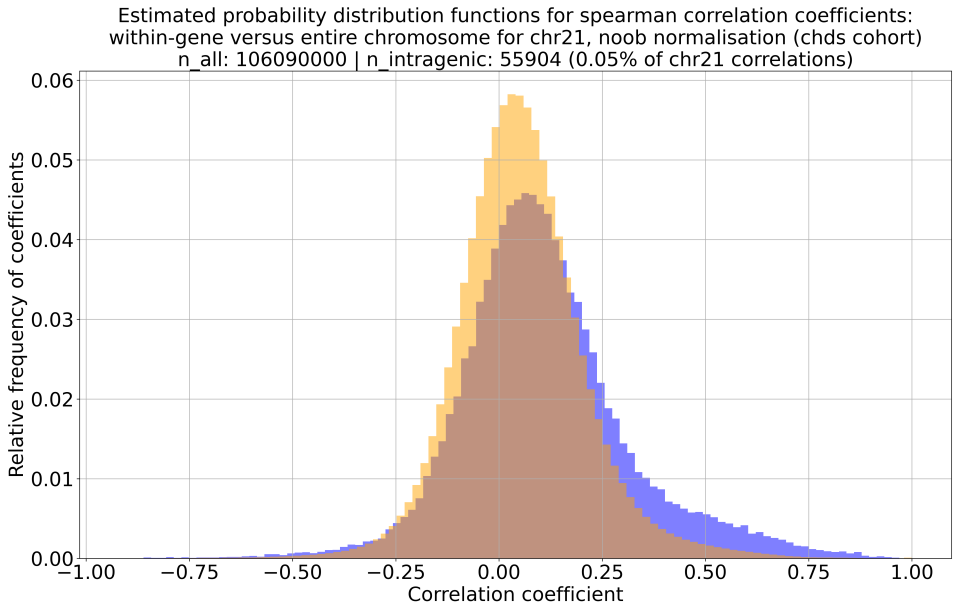

Figure 7.23: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 21, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

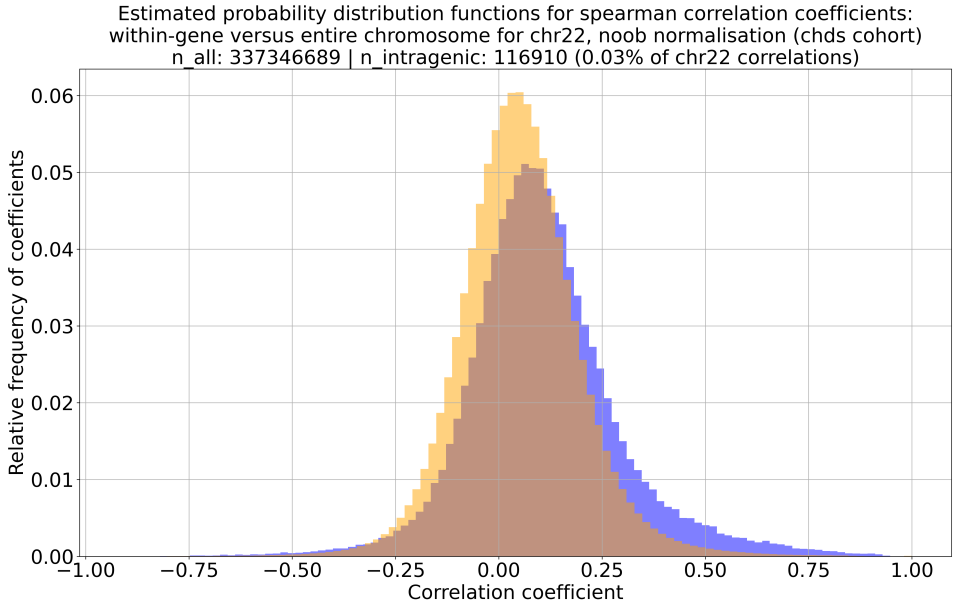

Figure 7.24: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome 22, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

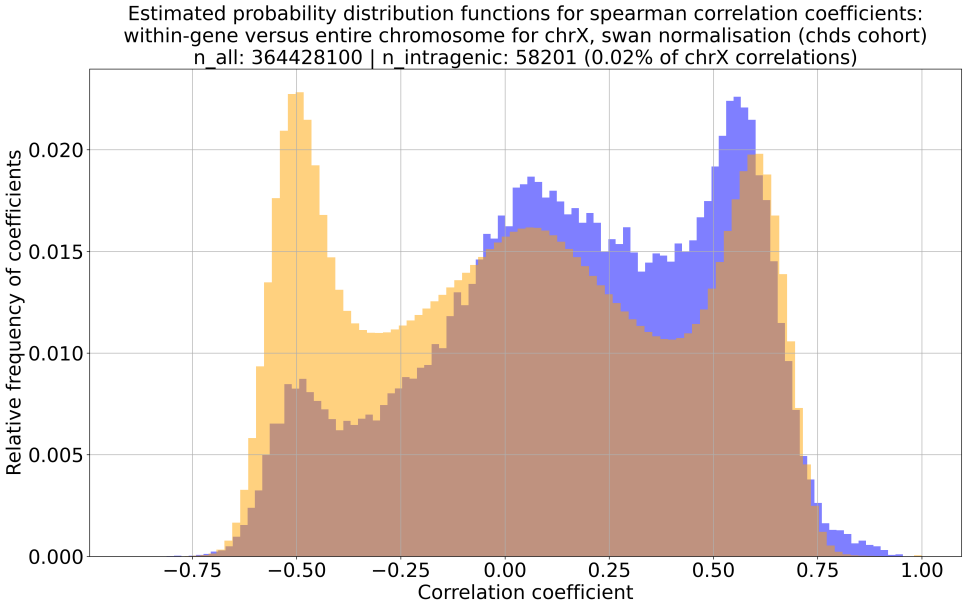

Figure 7.25: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome X, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

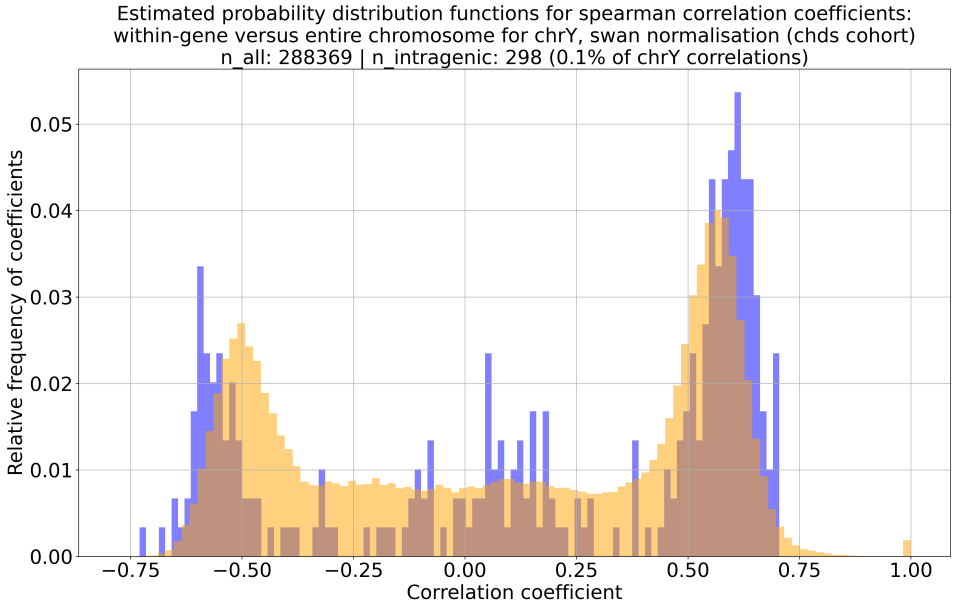

Figure 7.26: Frequency of correlation coefficients for whole genome (orange) and subset of correlating pairs from CpGs within the same gene (blue) for chromosome Y, excluding all CpG sites on annotated CpG islands. Source data: CHDS dataset

In this section, we see similar results to what we saw in section 7.2.2.1. Visual comparison suggests that the distribution of same-gene-excluding-island correlations is biased more to the right (positive correlations)

than same-gene correlations that include CpG islands. Otherwise, much of the commentary is the same.

7.3 Study: Correlation trends within pathways

This study aims to find evidence of strong correlation in methylation intensity between CpG sites on different genes on the same biological pathway.

7.3.1 Methods

We used the same methods from section 7.2.1 to select chromosomes, obtain correlation matrices and convert CpG sites to genes. We search the reactome database for biological pathway information, using techniques described in section 2.3.

The mean positive and negative correlations are calculated for the following subsets of each chromosome:

- CpG pairs on genes that share a pathway as detected by our automated search tool, except for those that are on the same gene
- All CpG pairs within that chromosome

Correlations within the same gene are investigated in section 7.2. To remove the effects due to correlations within the same gene, all correlations within the same gene would be removed from the subset based on genes with common pathways; this leaves only correlating pairs with CpG sites on separate genes, within the same pathway. The other subset represents the chromosome as a whole - it will include genes and pathways that were not correctly identified by our automated search tool as well as anything that has yet to been added to the database that we used.

To investigate the significance of the difference between the two subsets, we use a one-way ANOVA test using the Python *scipy* module. ANOVA tests are conducted separately for the positive and negative correlations.

7.3.2 Results

	Whole chromosome	Shared pathways
Number of CpG sites	36771	2426
Number of known genes assessed	440	440
Identified overlapping pathways	189	189
Number of positive correlating pairs	403073784	1532486
Number of negative correlating pairs	272961051	1179790
Mean positive correlation	0.146	0.152
Mean negative correlation	-0.103	-0.111
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	3630.0	9060.0
One-way ANOVA p-value	0.0	0.0

Table 7.1: Statistical comparison of correlations between genes on shared pathways on chromosome 4, and all correlations within chromosome 4 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	38452	2248
Number of known genes assessed	391	391
Identified overlapping pathways	183	183
Number of positive correlating pairs	449077532	1545450
Number of negative correlating pairs	290181394	1115880
Mean positive correlation	0.139	0.136
Mean negative correlation	-0.0986	-0.103
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	1330.0	2470.0
One-way ANOVA p-value	1.44e-291	0.0

Table 7.2: Statistical comparison of correlations between genes on shared pathways on chromosome 8, and all correlations within chromosome 8 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	26167	2407
Number of known genes assessed	428	428
Identified overlapping pathways	189	189
Number of positive correlating pairs	210811274	1475228
Number of negative correlating pairs	131531587	973470
Mean positive correlation	0.133	0.135
Mean negative correlation	-0.0953	-0.0979
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	711.0	918.0
One-way ANOVA p-value	1.19e-156	1.44e-201

Table 7.3: Statistical comparison of correlations between genes on shared pathways on chromosome 9, and all correlations within chromosome 9 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	21040	1236
Number of known genes assessed	196	196
Identified overlapping pathways	81	81
Number of positive correlating pairs	136109854	794354
Number of negative correlating pairs	85220426	518162
Mean positive correlation	0.142	0.139
Mean negative correlation	-0.0997	-0.1
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	358.0	28.0
One-way ANOVA p-value	6.91e-80	1.21e-07

Table 7.4: Statistical comparison of correlations between genes on shared pathways on chromosome 13, and all correlations within chromosome 13 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	29550	2033
Number of known genes assessed	343	343
Identified overlapping pathways	165	165
Number of positive correlating pairs	266185214	2723202
Number of negative correlating pairs	170401261	1892004
Mean positive correlation	0.135	0.135
Mean negative correlation	-0.0992	-0.105
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	0.000914	7450.0
One-way ANOVA p-value	0.976	0.0

Table 7.5: Statistical comparison of correlations between genes on shared pathways on chromosome 14, and all correlations within chromosome 14 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	28741	2357
Number of known genes assessed	364	364
Identified overlapping pathways	150	150
Number of positive correlating pairs	254297067	1492116
Number of negative correlating pairs	158711103	968528
Mean positive correlation	0.135	0.138
Mean negative correlation	-0.0971	-0.1
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	1200.0	1350.0
One-way ANOVA p-value	5.12e-263	4.75e-295

Table 7.6: Statistical comparison of correlations between genes on shared pathways on chromosome 15, and all correlations within chromosome 15 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	37939	2415
Number of known genes assessed	450	450
Identified overlapping pathways	190	190
Number of positive correlating pairs	449528374	2844484
Number of negative correlating pairs	270136517	1848404
Mean positive correlation	0.125	0.126
Mean negative correlation	-0.0928	-0.0963
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	167.0	3270.0
One-way ANOVA p-value	3.41e-38	0.0

Table 7.7: Statistical comparison of correlations between genes on shared pathways on chromosome 16, and all correlations within chromosome 16 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	14899	810
Number of known genes assessed	134	134
Identified overlapping pathways	73	73
Number of positive correlating pairs	68742921	280860
Number of negative correlating pairs	42239730	190970
Mean positive correlation	0.143	0.14
Mean negative correlation	-0.0995	-0.103
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	136.0	263.0
One-way ANOVA p-value	2.29e-31	3.8e-59

Table 7.8: Statistical comparison of correlations between genes on shared pathways on chromosome 18, and all correlations within chromosome 18 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	22960	1734
Number of known genes assessed	294	294
Identified overlapping pathways	152	152
Number of positive correlating pairs	165890096	1043752
Number of negative correlating pairs	97679224	664018
Mean positive correlation	0.131	0.131
Mean negative correlation	-0.0936	-0.0964
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	4.88	768.0
One-way ANOVA p-value	0.0272	3.85e-169

Table 7.9: Statistical comparison of correlations between genes on shared pathways on chromosome 20, and all correlations within chromosome 20 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	10300	931
Number of known genes assessed	127	127
Identified overlapping pathways	38	38
Number of positive correlating pairs	33906857	183758
Number of negative correlating pairs	19132993	91590
Mean positive correlation	0.137	0.144
Mean negative correlation	-0.0951	-0.0946
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	797.0	2.74
One-way ANOVA p-value	2e-175	0.0978

Table 7.10: Statistical comparison of correlations between genes on shared pathways on chromosome 21, and all correlations within chromosome 21 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	18367	1367
Number of known genes assessed	235	235
Identified overlapping pathways	126	126
Number of positive correlating pairs	108741296	740188
Number of negative correlating pairs	59922865	463874
Mean positive correlation	0.13	0.129
Mean negative correlation	-0.0921	-0.101
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	10.5	4790.0
One-way ANOVA p-value	0.00121	0.0

Table 7.11: Statistical comparison of correlations between genes on shared pathways on chromosome 22, and all correlations within chromosome 22 in general, from correlation matrices derived from noob-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	19090	1895
Number of known genes assessed	343	343
Identified overlapping pathways	140	140
Number of positive correlating pairs	97951388	609790
Number of negative correlating pairs	84253117	480862
Mean positive correlation	0.361	0.389
Mean negative correlation	-0.322	-0.327
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	9680.0	312.0
One-way ANOVA p-value	0.0	8.29e-70

Table 7.12: Statistical comparison of correlations between genes on shared pathways on chromosome x, and all correlations within chromosome x in general, from correlation matrices derived from swan-normalised beta values (cohort: chds)

	Whole chromosome	Shared pathways
Number of CpG sites	537	36
Number of known genes assessed	6	6
Identified overlapping pathways	3	3
Number of positive correlating pairs	83330	238
Number of negative correlating pairs	60586	186
Mean positive correlation	0.442	0.557
Mean negative correlation	-0.371	-0.511
	Positive correlations	**Negative correlations**
One-way ANOVA F-score	89.5	121.0
One-way ANOVA p-value	3.12e-21	3.27e-28

Table 7.13: Statistical comparison of correlations between genes on shared pathways on chromosome y, and all correlations within chromosome y in general, from correlation matrices derived from swan-normalised beta values (cohort: chds)

For all autosomes, there is no regular difference in mean positive and negative correlation. ANOVA scores generally had quite low P-values, so differences were likely present; but this lack of consistency means we can not make assertions on the relationship between correlation strength and presence of a common pathway between CpG sites for autosomes. For allosomes however, the average magnitude of positive and negative correlations is higher in both cases. This was validated by ANOVA, with $p << 0.001$ in all cases. It should be considered that the number of overlapping pathways on the Y chromosome is very small compared to most other chromosomes (typically two orders of magnitude fewer) so the dataset we are working with here is more prone to the statistical effects experienced by small datasets.

Chapter 8

General Discussion

8.1 Overview

In this thesis, we sought to find epigenetic trends by looking at the correlation between DNA methylation intensities of CpG sites within the human genome, from data derived from microarray analysis of whole blood samples. Our studies have found evidence to support several of our hypotheses. Some key findings include:

- Choice of array normalisation method has a significant impact on the ability to detect underlying associations in DNA methylation data
- Methylation intensities of CpG sites within CpG islands tend to correlate more positively than the average for a chromosome
- Correlations tend to be stronger between CpG sites associated with the same gene, and this effect isn't entirely due to CpG islands

These findings have made use of largely-ignored or unconsidered opportunities in DNA methylation correlation analysis, and are elaborated upon in later sections of this chapter.

More generally, the success we've had with our novel approach is encouraging. Throughout this thesis, we have discussed some of the implications of DNA methylation and why the study thereof is a valuable and medically-important endeavour. Our correlation studies have been, in the grand scheme of things, very limited in scope; can we make use of similar methods in future studies? As more data becomes available and our processing capabilities improve, we will have an abundance of opportunity for further studies in this area. Some potential avenues for future research are touched upon in section 8.5, though the purview of what we can do with DNA methylation studies is significantly more broad.

8.2 Technical discussions

Several studies earlier in this thesis had outcomes that influenced how the studies in part III. Rather than include them in this discussion, they have been incorporated into their corresponding chapters as the context of each study greatly improves readability for the discussion. The reader is reminded of the following discussions:

- Section 3.2.2: A preliminary study of the statistical properties of beta values for chromosome 21
- Section 3.3.2: A preliminary study of the statistical properties of beta correlation matrices for chromosome 21
- Section 3.3.2.3: A preliminary network analysis of a beta correlation matrix for chromosome 21
- Section 4.6: A discussion of the impact of normalisation type on beta correlation matrices
- Section 5.4: A discussion of the impact of correlation method on beta correlation matrices

8.3 Implications of results from biological studies

8.3.1 The methylation intensities of CpG sites within CpG islands tend to correlate more positively than the average for a chromosome

The results of section 6.4 showed that a number of the strongest 10% of positive correlations were present along the diagonal of the beta correlation matrix. These are not self-correlations; section 2.2.5 explains that these are removed. Rather, they are correlations between the methylation intensity of two CpG sites in relative proximity to each other. It was hypothesised that these strong correlations were on CpG islands (which are discussed in section 1.2), and that these islands may contain evident and biologically-relevant correlations owing to their role in gene regulation. This study was followed up by another study (in section 6.5) that explicitly investigated correlations within CpG islands.

The results in section 6.5 showed that, without exception, all chromosomes tested had a higher average positive correlation within CpG islands than the baseline level of the entire chromosome. Similarly, it showed that adjacent (in the available data) CpG pairs had a higher average positive correlation than the already-elevated average of the CpG island subset. In both cases, the ANOVA results suggested that these results were statistically significant ($p << 0.001$).

A potential explanation for this phenomenon is that CpG islands, being regulatory elements, tend to be consistently hypo- or hyper-methylated (Bird 1986; Esteller 2002) and this would manifest as a higher correlation score between CpG sites within the same island.

CpG islands are only loosely defined and their definition has always been somewhat objective. UCSC use the following algorithm for prediction of CpG islands (University of California Santa Cruz, 2020):

- Each dinucleotide within the genome is scored (+17 for CG and -1 for others) and maximally-scoring segments are identified (it is not explained how). Each segment is evaluated for the following criteria:
  - GC content of 50% or greater
  - length greater than 200 bp
  - ratio greater than 0.6 of observed number of CG dinucleotides to the expected number on the basis of the number of Gs and Cs in the segment
- The CpG count is the number of CG dinucleotides in the island. The Percentage CpG is the ratio of CpG nucleotide bases (twice the CpG count) to the length. The ratio of observed to expected CpG is

calculated according to the formula $\frac{Obs\,CpG}{Exp\,CpG} = \frac{Number\ of\ CpG \times N}{Number\ of\ C \times Number\ of\ G}$ where N is the length of the sequence (as per the formula cited in Gardiner-Garden and Frommer, 1987).

Our findings suggest that it may be possible to use strong correlations to identify new CpG islands or confirm existing ones. This is discussed further in section 8.5.1.

Average negative correlations tended to be more positive for both tested subsets, for most chromosomes, but the relatively high p-values in several cases as well as the inconsistency of trend means that we have insufficient evidence to suggest anything regarding this.

8.3.2 There is insufficient evidence to suggest a significant linear relationship between distance and correlation strength

Section 6.2 used linear regression to investigate the relationship between the distance between two CpG sites on the same chromosome, and their degree of correlation. A similar investigation into the same relationship between CpG sites on the same genes was undertaken in section 6.3 and similar results were obtained. For all chromosomes, the R-squared value of the regression is at or near zero. R-squared can be thought of as the proportion of variation explained by the model, with an R-squared close to one meaning that more of the variation in the data is explained by the model, and vice versa. Across the entire chromosome, all of our linear regression models all have an extremely low R-squared, never exceeding 0.01 (which occurred in the Y chromosome) - this suggests that their predictive capability is practically negligible. The same was largely shown for correlations within the same gene, though the sex chromosomes possessed an R-squared of roughly 0.02 for positive correlations, and the relationship between same-gene negative correlations and distance on the Y chromosome was roughly 0.09. We deem these R-squared values too low for their models to provide any meaningful insights, and therefore we do not consider there to be sufficient evidence to suggest a relationship between distance and correlation strength.

Another factor that suggests no significant linear relationship between distance and correlation strength is the tenuousness of the gradients of each linear regression as shown in the tables of the relevant results sections. We would expect a general trend to hold in terms of direction of gradient for positive and negative correlations; e.g. assuming correlation strength was proportional to distance, all regressions of positive correlations would have a negative gradient, and all regressions of negative correlations would have a positive gradient. This is not the case. We can see multiple examples of both positive and negative linear regressions having gradients in the same direction, and this direction is not consistent within these examples.

A key limitation to our methods could be the use of linear models, rather than taking a more flexible approach to modelling. Rather than try out a number of different models, it was decided that the best course of action was taking a different approach and looking and correlations within functional gene groups, which tend to be co-located on the same chromosome (Thévenin et al. 2014). Studies investigating this include those in sections 6.5, 7.2 and 7.3, discussions of which are included in this chapter.

8.3.3 Correlations tend to be stronger within genes

Despite the apparent lack of relationship between distance and strength described in section 8.3.2, our histograms in section 7.2.2.1 suggest that the statistical distribution of correlation strengths is weighted more

towards the strong positive side for all tested chromosomes.

This was initially hypothesised to be caused by CpG islands. The presence of CpG islands within genes has long been studied (Larsen et al. 1992) and we discussed in section 8.3.1 that CpG islands tend to correlate more strongly; consequently, inclusion of correlations between sites on CpG islands would shift the histogram to the right. To test this hypothesis, we generated another set of histograms that did not include correlations involving CpG sites associated with CpG islands. These are shown in section 7.2.2.2. Assuming that our hypothesis was correct, the distribution should have shifted back towards the mean (perhaps even following the distribution of the overall chromosome if the islands were the only factor), but what we saw (in autosomes) was the opposite - rather than shift towards the mean, the distribution of correlations actually became more skewed to the right, suggesting they tended to become stronger when we didn't consider the CpG islands. This suggests that correlations in methylation intensity between CpGs within the same genes are stronger on average, in comparison to CpGs across the chromosome as a whole, and this effect occurs outside CpG islands, at least in autosomes. We did not see the same trend in allosomes, though it is difficult to draw conclusions from their histograms as they follow more-complex distributions.

One possible explanation could be that the regulatory effect of CpG islands is dispersed throughout the gene, so even if a CpG site is not considered part of an existing CpG island (or a functional but unannotated island) it could still have similar correlations between the methylation intensities of its CpG sites. More generally, given that we know that gene regulation is highly complex and involves more than methylation at CpG islands, other regulatory mechanisms could well be acting upon the gene to ensure that transcription is activated/repressed consistently.

8.3.4 There is insufficient evidence to suggest a significant relationship between correlation strength and presence of a common pathway, for autosomes

Section 7.3 compared the average positive and negative correlation between all chromosomes, with results available in section 7.3.2. It must be acknowledged that we are limited in what pathways we are able to assess with our data. There are multiple limiting factors here (described more generally in section 8.4):

- We only data for CpG sites probed in the EPIC array
- The database we used for pathway information doesn't yet contain data for all pathways across the human genome
- The automated search tool, being an unsupervised algorithm, may erroneously detect pathways or gene products where it shouldn't, or miss those that it should

With the above in mind, we still had a substantial subset of data to work with, with generally 5-10% of all CpG sites within a chromosome situated on shared pathways. Results were generally not in agreement with hypotheses. For all autosomes, there is no regular difference in mean positive and negative correlation. ANOVA scores generally had quite low P-values, so differences were likely present; but this lack of consistency means we can not make assertions on the relationship between correlation strength and presence of a common pathway between CpG sites for autosomes. We discuss allosomal results in section 8.3.5.

It should be pointed out that in table 7.5 we can see a p-value of 0.976 for the one-way ANOVA test for positive correlations, which is high relative to that of other chromosomes. This will most likely be due to the fact that both tested subsets have the same mean.

8.3.5 Correlations to be stronger between genes on shared pathways within the sex chromosomes

As discussed in section 8.3.4, there was no evidence to suggest a significant relationship between correlation strength and the presence of a common pathway for non-sex chromosomes. The same study showed that for the X and Y chromosomes, the average magnitude of positive and negative correlations is higher in both cases. This was validated by ANOVA, with $p << 0.001$ in all cases.

For the X chromosome, it is proposed that many of the strong correlations arise due to co-methylation occurring as part of X-inactivation (discussed in section 1.2.3). This phenomenon 'switches off' an entire chromosome, and as we are looking at pathways on the same chromosome, they will all be switched on or off together, manifesting as a notably stronger positive correlation - though this 'signal' will apply to all CpG sites on the chromosome, so it is not pathway-specific. A potential explanation to why correlations between pathways tends to be stronger on the X chromosome is dosage compensation, an epigenetic phenomenon wherein the level of transcription is altered in some X-linked genes (Lucchesi et al. 2005). It should be pointed out that Lucchesi et al. explain that X-inactivation is one mechanism of dosage compensation, but others exist.

For the Y chromosome, it is possible that statistical effects due to small numbers of genes on the same pathway may make these results significant, despite no underlying biological phenomenon. Future studies could make use of new information regarding genes and pathways on the Y chromosome to validate the results of this study.

8.3.6 There are regions within chromosomes wherein CpG sites appear to have a significantly increased tendency to strongly correlate in methylation intensity with other CpG sites

In section 6.4, we show the locations of pairs CpG sites which correlate strongly in methylation intensity. Our plots show a distinct banded pattern in all cases.

One possibility is that these bands are biologically-relevant and are subjected to epigenetic regulatory pressures. They could be regulatory regions that all tend to be 'switched on or off' in response to a common environmental influence - this possibility was described in section 1.3.3. An alternative explanation could be that homology between these regions results in similar activity at these regions by active methylation or demethylation enzymes. This was not explored in studies in this thesis, but is a potential avenue for future research.

It should be pointed out that we can also see regions where no strong correlations occur in our data. These could be regions of relatively low methylation activity; it could be that their regulation is dominated

by mechanisms other than DNA methylation, such as histone modification. Studies have shown that DNA methylation and histone modification in particular are linked (e.g. reviewed in Cedar and Bergman, 2009) though further research would be required to determine if altered correlations in methylation intensity occur due to other epigenetic mechanisms.

A more mundane explanation for the described findings could be technical bias in the underlying array. While we attempt to account for background fluorescence, our methods can never truly eliminate it; this is simply a limitation of the technology we use to capture DNA methylation information. It could be that some probes (or clusters of them on the array) are more prone to background fluorescence than others. Perhaps another issue could be 'crosstalk' between probes during data acquisition - this may be too faint or difficult to identify in beta values, but if it were consistent between samples, it would manifest as an increased correlation coefficient. Further technical studies into the phenomena affecting the EPIC array would be needed to confirm either of these explanations.

8.4 General limitations

8.4.1 Microarray data is limited in scope

Our analyses have a number of limitations. The first (which we consider to be most significant) is that we used data from the EPIC array for our analysis, which only looks at a small fraction of CpG sites across the genome. This severely limits the our visibility of the general methylome. Selection bias may also be a significant factor as probes for the EPIC array were likely chosen based on some technical criteria rather than being randomly selected from the genome as a whole. The EPIC array can be used to investigate roughly 3% of the human genome, as discussed in section 2.1.1. We would need to use other techniques, such as whole-genome bisulfite sequencing, to increase the number of CpGs we have available for analysis.

8.4.2 Cohort size

Most of our analyses only used a single cohort of 120 individuals. We do not know if this is a sufficient number for calculation of 'stable' correlations, in the context of DNA methylation studies. A 2013 study by Schönbrodt and Perugini (corrected in 2018) suggest that the point of stability for correlations is in the vicinity of n = 161; the CHDS cohort has about 75% of this. The Monte Carlo simulation used by Schönbrodt and Perugini for their study used bootstrapped samples from a bivariate Gaussian distribution; this may not translate particularly well to the distribution of methylation intensities at a CpG site for a given cohort due as the presence of SNPs and other confounding factors. The future potential for research into a minimum cohort size for stable correlation calculation is discussed in section 8.5.7.

8.4.3 Incomplete annotation of genes

As of early 2021, we are still facing a number of unresolved gaps and issues with our most modern and well-studied human genome assemblies (Genome Reference Consortium, 2021). This, combined with the fact that researchers are still at odds over the number of genes in the human genome, means that our existing set of gene annotations are likely to remain incomplete for some time. Our research can only make use the existing subset of annotations so our studies related to correlations within or pertaining to genes do not use all of them in the entire human genome.

8.4.4 Incomplete pathway databases

Researchers around the world have been contributing to our knowledge of biological pathways, but given the relative infancy of bioinformatics in the grand scheme of things, we are a long way off being able to say with certainty that we have a complete database of every pathway in the human genome. The Reactome database that we used in studies for this thesis is particularly well-used, with publications detailing the database being released every few years, and thousands of other papers citing these. Reactome has provided a great amount of information for this thesis but is ultimately a long way from being a comprehensive source of every pathway that we could consider in our research. Additionally, further studies across a number of fields may result in us revising our understanding on pathways that had been supported by previous research; future updates to the database may edit or remove existing ones, so our results come with the caveat that they're based off information that is currently accurate, but may not be in future.

8.5 Future opportunities

8.5.1 Using correlations to identify CpG islands

We discussed in section 8.3.1 how the methylation intensity of CpG sites within CpG islands (as arbitrated by UCSC) tend to correlate more positively with each other, compared with correlations across the chromosome as a whole. We discussed how the regulatory nature of CpG islands may be the cause of these higher-than-average correlations. Assuming this is the case, correlation could be used as an alternative measure to either confirm existing islands or possibly extend/shorten/subdivide them based on functional CpG groups. Currently, CpG islands are identified using genomic methods, such as using the frequency of CpG dinucleotides in a subsequence as an indicator (Gardiner-Garden and Frommer, 1987). I propose that we make use of epigenetic data to delineate CpG islands, as much of their biological relevance appears to be in the context of epigenetic regulation. Improving our understanding and demarcation of these regulatory areas may have applications for the design of DNA methylation microarrays (for example), as we would be able to make more-informed selections of which CpG sites to add to these arrays.

8.5.2 Improving our ability to identify strong correlations in DNA methylation data

One of the limitations discussed in section 1.3.1.3 was that many of our strong correlations could be due to random chance. Further research could examine methods of accounting for randomness or identify different approaches which could be used when looking for biologically-relevant correlations. This would greatly enhance the potential to extract meaningful insights from correlation studies as our findings would come with fewer caveats.

8.5.3 Correlation analysis using whole-genome data

As discussed in section 8.4, use of the EPIC array limits our analyses to a small proportion of the methylome. DNA methylation datasets obtained via whole-genome bisulfite sequencing have been produced for years and the recent advent of techniques to measure DNA methylation with nanopore technologies will improve the ability for more full-methylome datasets to be made. Both of these techniques are described in section 2.1.1. There will certainly be trends in the $\sim 97\%$ of the methylome that we were not able to access in studies in

this thesis, so analysis of data obtained by more comprehensive techniques is a significant future opportunity. In particular, a full-methylome method would provide much greater resolution when used to identify CpG islands with correlations.

8.5.4 Investigating the effects of aging on correlations

An epigenetic phenomenon associated with aging is dysregulation of DNA methylation, as discussed in section 1.2.6. The primary dataset used in this thesis, CHDS, consisted of 120 individuals whose age was within a year of 28 at time of data collection. A comparable dataset of 120 individuals at later ages was not available for this study; while several informal investigations undertaken alongside several studies in this thesis looked at the differences that may have been caused by age, there was insufficient data to make robust scientific claims. A future study could look into the effects of age on correlation in more detail, provided the dataset was sufficient. An example of a sufficient dataset would be a future assessment of the CHDS cohort, so the methylomes of individuals can be tracked at a range of different ages. This would also ensure that genetic and environmental effects were relatively consistent.

One observation I would hypothesise is that the average correlation between methylation intensities would decrease as one ages. There are several reasons for this:

- A decrease in epigenetic regulation weakens the 'epigenetic signal', which would result in weaker associations in DNA methylation and consequently lower correlation coefficients
- We have seen that there is generally a bi-modal distribution in DNA methylation values across the genome (e.g. in section 3.2.2) - the global decrease in methylation associated with age (Bollati et al. 2009; Heyn et al. 2012) would flatten and shift the right-most peak of the distribution to the left, increasing variance and therefore 'noise' within that peak which may obscure the monotonic changes that we hope to find using the Spearman rank correlation method

Another hypothesis is that different CpG sites correlate at different life stages. For example, a cell from an embryo undergoing early development may have different epigenetic associations than an individual who is in later stages of life.

Research into both of these hypotheses could yield benefits for the study of aging, such as identifying which pathways remain consistent into old age, which ones are the first to deteriorate, etc. This research could have applications in treatment of the diseases associated with aging, which will only grow more frequent as the average age of our population increases.

8.5.5 Investigating the effects of tissue type on correlations

The influence of tissue type on DNA methylation was discussed briefly in section 1.2.5. For the purposes of this study, we have only have whole blood samples available from the CHDS cohort. This has been acceptable for a proof of concept, but there is significant potential for research into correlations in DNA methylation in different tissue types. I hypothesise that there will be some strong correlations that are present in most tissue types (perhaps even all of them) as they are simply required for continued survival of the cell. I suggest that the overlap of strong correlations is higher for similar cell types (e.g. in skin and alveoli, both squamous epithelial cells) than for those in different cell types (e.g. skin and brain tissues).

8.5.6 Investigating correlations using multiple samples from a single individual

All studies in this thesis used a single sample with multiple individuals. Something that may provide scientific value would be a study of multiple samples from the same individual. It is proposed that correlations within the methylome are influenced by epigenetic phenomena. A dysregulation of epigenetically-grounded regulatory systems could lead to a substantially different correlation profile for specific CpG sites, between different samples. In general terms, a given set of CpG sites may be expected to correlate with a certain strength, and deviation from this could suggest a disease phenotype or epigenetic dysfunction. This could occur spatially or temporally:

- Different correlation profiles between samples from the same tissue type could indicate disease, e.g. there may be a set of CpGs which correlate less in methylation intensity in cancerous or pre-cancerous conditions in the skin, compared with healthy skin
- Different correlation profiles between repeated samples from the same tissue type over time could also indicate disease progression, e.g. if a consistent pattern of correlation in some set of CpGs is observed for a long period of time (on the scale of years or decades), a sudden change may suggest the onset of a disease phenotype.

Genetic factors would obviously dominate any epigenetic signal if correlations were calculated from tissue samples within a single individual, so a significant amount of research would be required to identify which genetic variants can alter epigenetic associations.

8.5.7 Identification of a minimum cohort size for stable calculation of correlations

We discussed briefly in section 8.4.2 that there exists a minimum size for calculation of stable correlation coefficients. One of the general issues with correlations (and indeed quantitative statistics as a whole) is that cohorts only represent a small sample of the available population and many factors contribute to 'noise' in the data which adversely affects the similarity of a metric calculated for a sample (correlation in this case) to that overall metric for the population. Correlation coefficients in particular are prone to this noise as they rely on two noisy data points, so sample-based effects can be quite pronounced in the data. The primary cohort used in this thesis had about 75% of the value recommended in a past Monte Carlo simulation (Schönbrodt and Perugini, 2013) - this was the largest and best-curated cohort available for these studies, so was the best option despite having a smaller sample size that this recommendation. That said, the recommended value was based on normally-distributed pairs, and the preliminary study in chapter 3 suggested that the distribution of average beta values is more complex than that - the effects on minimum cohort size due to this distribution, as well as biological phenomena such as SNPs and technical aspects such as normalisation type, are yet to be seen.

Regardless, there will be a minimum cohort size required for stable calculation of correlations. With more data on the distribution of methylation intensity across each CpG site, it may be possible to conduct a Monte Carlo simulation similar to that undertaken by Schönbrodt and Perugini, from which we can determine the minimum number of individual samples needed for future correlation studies.

8.5.8 Correlation analysis of non-5mC methylation

Our research, and indeed much of the worldwide research into DNA methylation, has focused specifically on 5-methylcytosine as the primary mark of methylation. Other bases are known to undergo methylation as well - for example, N^6-methyladenine conversion was described briefly in section 1.2. As part of the ongoing efforts to understand epigenetic tags other than 5mC, we could conduct correlation analyses of m6A. This could involve reassessment of hypotheses of earlier studies in the context of adenine methylation or investigation of some of the potential future avenues of research described in this section.

8.5.9 Development of combined cohorts

Researchers develop their own cohorts based on the requirements of their own research and these are often suitable for further analysis in other studies. This is not always the case, however. Smaller research cohorts may have insufficient subjects to derive statistically-significant information (perhaps having as few as a single subject) and if they are to be used in a larger EWAS, they must be combined with other cohorts. This poses a number of challenges owing to the impact of factors discussed in 1.2). Nonetheless, we can still make efforts to combine cohorts for use in broader studies, provided we are aware of the caveats.

We expect the state of DNA methylation to be highly variable between different individuals. A very simplified model for the methylation intensity of each CpG site in the genome for a specific tissue type in any arbitrary individual could be defined as:

$$x(t) = \bar{x}(t) + g(t) + e(t) + i \tag{8.1}$$

where $x(t)$ is a vector of methylation intensities of each CpG site (as a function of time/chronological age), $\bar{x}(t)$ is the 'baseline' methylome for any arbitrary human at a given age, $g(t)$ is the genetic influence, $e(t)$ is the environmental impact specific to that individual, and i is the one-off impact of inheritance (with long-term inherited factors considered as part of $g(t)$ and $e(t)$). Of these quantities, only $\bar{x}(t)$ is not specific to the individual - rather, this is something that we can estimate. The other factors are all specific to the individual. Both $g(t)$ and $e(t)$ would be dependent on the individual's genome and epigenetic landscape for that tissue type; both of these dependencies change over time (albeit less so in case of the genome), and the epigenetic landscape also has genomic dependencies, so there's a great deal of complexity if we consider them separately. Instead, if we combine the individual-specific factors into a single drift factor $d(t)$, we can define our estimate for the average methylation intensity for a given tissue type within an individual as:

$$\hat{x}(t) = \bar{x}(t) + d(t) \tag{8.2}$$

Further to this, if we assume the drift factor to be stochastic with a mean of zero, then for a sufficiently large cohort, the average drift factor will tend towards zero and the average of all individual methylation intensities will be approximately equal to the baseline epigenetic landscape.

$$\hat{x}_{cohort}(t) \approx \bar{x}(t) \text{ for a sufficiently large cohort} \tag{8.3}$$

The larger the cohort, the better the cohort will be at representing of the human species in general. Selection bias is an unavoidable issue but we can still get a pretty good idea of the epigenetic characteristics of a large

number of CpG sites and their associated genes, especially if their pathways tend to be conserved between people of different genetic heritage.

A key practical difficulty of combining cohorts for EWAS is that we do not yet have the ability to perfectly assess the methylation intensity of any particular CpG site within the human body. Any measurement of this is going to be prone to some level of error, causing results to be less accurate than we'd like. This error is caused by factors specific to the technologies being used. We can define the estimate of methylation intensity $\hat{x}(t)$, taken at a specific time t_s for a tissue type in a given individual as:

$$\hat{x}(t_s) = x(t_s) + \epsilon_f \tag{8.4}$$

Where $x(t_s)$ is the actual methylation intensity. The error term ϵ_f comprises technology-specific issues such as background fluorescence (in the case of arrays) and sequencing issues (e.g. for WGBS). We can make the assumption that ϵ_f is stochastic and can be modelled as a normal distribution with a mean of μ_f and standard deviation of σ_f. In practice, ϵ_f is unlikely to follow a perfect normal distribution, but for the sake of discussion, we can model ϵ_f as:

$$\epsilon_f \sim \mathcal{N}(\mu_f, {\sigma_f}^2) \tag{8.5}$$

Each normalisation type will have a different μ_f and σ_f. For example, a set of unnormalised values for methylation intensity would be influenced by background fluorescence much more than values that were normalised, and this may lead to increased variance within a CpG site. This is studied in more depth in section 4.4. There may also be some dependency on CpG type as probe chemistry and presence of SNPs would have an effect on ϵ_f as well, and the effects of these may vary between the different normalisation types. Consequently, combining two datasets produced with different normalisation methods will result in a more complicated and thus less predictable error term which would adversely affect our ability to obtain a decent estimate of $\hat{x}(t)$.

A study where in cohorts are combined is beyond the scope of this thesis, though would certainly allow us to make better-informed choices when combining cohorts for new studies and meta-analyses. This discussion is presented to anyone undertaking such a study in the hopes that they may derive benefits from and expand upon it.

Appendix A
Computational resource configurations

Systems

System 1

CPU	Intel(R) Core(TM) i7-7700K CPU @ 4.20GHz
GPU	NVIDIA GeForce GTX 1060 6GB
RAM	32GB (2 x 16GB) DDR4 3200MHz
Data Drive	Western Digital Caviar Green 2TB (HDD)
System Drive	Samsung 850 EVO 250GB (SSD)
OS	Windows 10 Home

System 2

CPU	Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.10GHz x 16
GPU	N/A
RAM	Virtual (VMWare) 128GB
System/Data Drive	Virtual (VMWare) 6TB
OS	Ubuntu 20.04.1 LTS

System 2 was employed between November 2020 and March 2021 using the University of Canterbury Research Compute Cluster.

R

R Version: 3.6.1
IDE: RStudio 1.2.5001
BiocManager Version: 3.10
Minfi Version: 1.32
Profvis Version: 0.3.7

R was only used on system 1 for the purposes of preprocessing and normalising array data.

Python

Configuration 1

Python Version: 3.7.4
IDE: PyCharm 2019.2.1 Community Edition
numpy: 1.17.1
scipy: 1.3.1
pandas: 0.25.3
scikit-learn: 0.22
networkx: 2.4
matplotlib: 3.1.1
gtfparse: 1.2.1
statsmodels: 0.12.2

Python configuration 1 was used on system 1.

Configuration 2

Python Version: 3.8.5
IDE: N/A (all development used configuration 1) numpy: 1.19.4
scipy: 1.5.4
pandas: 1.15
scikit-learn: 0.23.2
networkx: 2.5
matplotlib: 3.3.3
gtfparse: 1.2.1
statsmodels: 0.12.2

Python configuration 2 was used on system 2.

Appendix B
Primary Cohorts

CHDS: Christchurch Health and Development Study

The Christchurch Health and Development Study is an ongoing longitudinal study of a cohort consisting of 1,265 individuals born in the Christchurch urban region in mid-1977. In this thesis, we use a subset of 120 individuals from this cohort, from which whole blood samples were collected at approximately age 28. These samples were subjected to methylation analysis using the Illumina Infinium MethylationEPIC array, providing data for over 850,000 CpG sites.

As of time of submission, further information regarding this cohort can be found at:
https://www.otago.ac.nz/christchurch/research/healthdevelopment/.

The manifest used for this dataset was the Infinium MethylationEPIC v1.0 B5, provided by Illumina.

MTAB-7069: Methylation arrays (MethylationEPIC) of longitudinal cord and peripheral blood from children aged 0, 5 and 10 years old

E-MTAB-7069 is a dataset made available by the European Bioinformatics Institute, containing data collected for a longitudinal study of epigenetic changes which occur over the first 10 years of life. There are 11 subjects in this study, with three blood samples from each subject, taken at 5 years of age, 10 years of age, and from the umbilical cord after birth.

As of time of submission, further information regarding this cohort can be found at:
https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7069/?array=A-GEOD-21145.

The original publication for this cohort is included in the bibliography (Pérez et al. 2019).

The manifest used to divide the dataset based on chromosome was the Infinium MethylationEPIC v1.0 B5, provided by Illumina.

Appendix C
Workflow

This appendix is included in case someone wants to re-use the scripts written for this thesis. They can currently (as of time of submission, March 26 2021) be found at https://gitlab.com/kjalaric/correlation_methylome_boreilly2021. Relevant commit hashes (ordered by timestamp) are:

- 12b50799391df9596b8304d56d3612838d3289a2 for the scripts and pipeline
- 17296aa3250ecb46abc38075b3d75464be6de94a for gene and pathway information scraped from the reactome database
- 6264faf32a4aea1ab5a4d816586315ec92fb0d20 for a very sparse readme
- 80b214eabed14f11ce4fc1cf4ddd9986bf21bbec for licensing information (this is also the final commit for the thesis)

Contact me if the repo is not available and I may be able to send you the files.

The repo also includes the full set of gene and pathway information scraped from the reactome database. As of time of submission, these can be found in pathway_jsons/ directory in the main repo. Those used for the project were added in the commit with hash 17296aa3250ecb46abc38075b3d75464be6de94a.

General workflow

An overview of methods can be found in section 2.2. This section covers how the main ones are implemented in practice.

File locations and directory structure are specified in `common/file_format.py`. A script to set up directories for a given cohort was written but has not been maintained in the later stages of the project. It can still provide a decent skeleton from which to work with though - `scripts/set_up_directories_for_new_cohort.py`.

Generating beta matrices

- Required: .idat files
- Produces: beta .csv file for a given normalisation type and cohort; chromosomal beta csv files for a given normalisation type, cohort and all chromosomes

I used minfi in R to process raw **.idat** files into beta matrices. The following R script shows an example of how to use minfi, using the CHDS cohort and NOOB preprocessing method:

```
library(minfi)
setwd("K:/source/chds")

# idat processing
samplesheet = read.metharray.sheet(getwd())
GRset = read.metharray.exp(targets = samplesheet) # force=TRUE as an argument if it doesn't work
GRset = preprocessNoob(GRset)
GRset = mapToGenome(GRset)
GRset = addSnpInfo(GRset) # possibly not needed

# get beta and write it as a csv file
beta = getBeta(GRset)
write.table(beta, "beta_noob.csv", sep=",", row.names=TRUE, col.names=NA)
```

These large files need to be split based on chromosome unless you have 6TB+ of RAM and an abundance of time. The script `pipeline/split_beta_into_chromosome.py` can split the file for you.

Generating and serialising a beta correlation matrix

- Required: chromosome beta .csv file for a given normalisation type and cohort
- Produces: correlation .pkl file for a given normalisation type, cohort, chromosome and correlation method

Beta correlation matrices are generated using *pandas* in Python. I serialise all of these into **.pkl** files as they're used frequently and generating them from scratch every time takes too long. An implementation using multiprocessing to speed thing up is in `pipeline/mp_correlation.py`. This needs to be configured based on how much RAM you want to use. Alternatively, `pipeline/correlation.py` can run one correlation procedure at a time.

By default, Spearman correlation matrices are generated. This is changed via an argument in the script.

Finding strong correlations

- Required: correlation .pkl file for a given normalisation type, cohort, chromosome and correlation method
- Produces: .txt file containing strong correlations (either proportional or thresholded)

Generating a list of strong correlations, using either the proportional or threshold method, can be done using the aptly-named `pipeline/find_strong_correlations.py`. This generates a .txt file (technically its a CSV with a different file extension, I might update this at some point but doing so during my thesis would break everything) containing the following information on each line:

- cpg1, cpg2, correlation coefficient

Generating beta matrix statistics

- Required: chromosome beta .csv file for a given normalisation type and cohort
- Produces: .json file of statistics for the supplied betas

`scripts/mean_versus_standard_deviation_betas.py` was used to generate statistical information and plots used in chapter 4.

`scripts/ratio_positive_to_negative_correlations.py` can be used to obtain the ratio of positive ($>= 0$) or negative (< 0) correlations in a given correlation .pkl.

Generating beta correlation matrix statistics

- Required: correlation .pkl file for a given normalisation type, cohort, chromosome and correlation method
- Produces: statistics .txt file for the supplied .pkl file, and (optionally) a .csv file containing statistics for each CpG site

Beta correlation matrix statistics can be generated using `pipeline/stats_for_correlation_df.py`. This outputs a general report with global statistics of the correlation matrix, as well as a **`.csv`** file with statistics for each individual CpG site within that chromosome.

An alternative implementation is `pipeline/stats_for_correlation_df_basic.py`. This version does not generate the .csv file or any statistics particular to a given CpG site (e.g. kurtosis within that site). In this thesis, it has been preferable to use this version for larger chromosomes as it is significantly more lightweight and the larger chromosomes generally are not used for the detailed studies that require CpG-specific information.

Finding all correlations within all genes on a chromosome

- Required: correlation .pkl file for a given normalisation type, cohort, chromosome and correlation method
- Produces: .csv file containing a list of all correlating CpG pairs that are on a same gene, where these CpGs are located, the strength of the correlation and an identifier for the database used to obtain the annotation

`pipeline/find_all_correlations_within_genes.py` provides a .csv file with the following scheme:

- gene, annotation source, cpg1, cpg1 mapinfo, cpg2, cpg2 mapinfo, distance between sites, correlation coefficient

The 'mapinfo' is the position of the CpG as a base pair locus, obtained from the Illumina manifest regardless of annotation source.

References

Amaratunga, D., & Cabrera, J. (2001). Outlier resistance, standardization, and modeling issues for DNA microarray data. In *Statistics in Genetics and in the Environmental Sciences* (pp. 17-26): Springer.

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet, 23*(2), 185-188. doi:10.1038/13810

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., & Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics (Oxford, England), 30*(10), 1363-1369. doi:10.1093/bioinformatics/btu049

Auger, N., Jugé, V., Nicaud, C., & Pivoteau, C. (2018). On the worst-case complexity of TimSort. *arXiv preprint arXiv:1805.08612*.

Bagni, C., Tassone, F., Neri, G., & Hagerman, R. (2012). Fragile X syndrome: causes, diagnosis, mechanisms, and therapeutics. *The Journal of clinical investigation, 122*(12), 4314-4322. doi:10.1172/JCI63141

Bandyopadhyay, A. K., Paul, S., Adak, S., & Giri, A. K. (2016). Reduced LINE-1 methylation is associated with arsenic-induced genotoxic stress in children. *BioMetals, 29*(4), 731-741. doi:10.1007/s10534-016-9950-4

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research, 21*(3), 381-395. doi:10.1038/cr.2011.22

Bartolomei, M. S. (2009). Genomic imprinting: employing and avoiding epigenetic processes. *Genes & development, 23*(18), 2124-2133. doi:10.1101/gad.1841409

Bartolomei, M. S., & Ferguson-Smith, A. C. (2011). Mammalian genomic imprinting. *Cold Spring Harbor perspectives in biology, 3*(7), a002592. doi:10.1101/cshperspect.a002592

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., . . . Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology, 28*(10), 1045-1048. doi:10.1038/nbt1010-1045

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., . . . Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics, 98*(4), 288-295. doi:https://doi.org/10.1016/j.ygeno.2011.07.007

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., & Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics, 1*(1), 177-200. doi:10.2217/epi.09.14

Bind, M.-A., Zanobetti, A., Gasparrini, A., Peters, A., Coull, B., Baccarelli, A., . . . Schwartz, J. (2014). Effects of temperature and relative humidity on DNA methylation. *Epidemiology (Cambridge, Mass.), 25*(4), 561-569. doi:10.1097/EDE.0000000000000120

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature, 321*(6067), 209-213. doi:10.1038/321209a0

Bollati, V., Schwartz, J., Wright, R., Litonjua, A., Tarantini, L., Suh, H., . . . Baccarelli, A. (2009). Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mech Ageing Dev, 130*(4), 234-239. doi:10.1016/j.mad.2008.12.003

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D., & Scandura, J. M. (2011). DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLOS ONE, 6*(1), e14524. doi:10.1371/journal.pone.0014524

Brockdorff, N., & Turner, B. M. (2015). Dosage compensation in mammals. *Cold Spring Harbor perspectives in biology, 7*(3), a019406-a019406. doi:10.1101/cshperspect.a019406

Cancer Research UK. (2020). Cancer incidence by age. Retrieved from https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age

Cao-Lei, L., Massart, R., Suderman, M. J., Machnes, Z., Elgbeili, G., Laplante, D. P., . . . King, S. (2014). DNA methylation signatures triggered by prenatal maternal stress exposure to a natural disaster: Project Ice Storm. *PLOS ONE, 9*(9), e107653. doi:10.1371/journal.pone.0107653

Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics, 10*(5), 295-304. doi:10.1038/nrg2540

Chahil, G., Yelam, A., & Bollu, P. C. (2018). Rett Syndrome in Males: A Case Report and Review of Literature. *Cureus, 10*(10), e3414-e3414. doi:10.7759/cureus.3414

Chahrour, M., Jung, S. Y., Shaw, C., Zhou, X., Wong, S. T. C., Qin, J., & Zoghbi, H. Y. (2008). MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science (New York, N.Y.), 320*(5880), 1224-1229. doi:10.1126/science.1153252

Chen, F., He, X., Luan, G., & Li, T. (2019). Role of DNA Methylation and Adenosine in Ketogenic Diet for Pharmacoresistant Epilepsy: Focus on Epileptogenesis and Associated Comorbidities. *Frontiers in neurology, 10*, 119-119. doi:10.3389/fneur.2019.00119

Cheng, Y., He, C., Wang, M., Ma, X., Mo, F., Yang, S., . . . Wei, X. (2019). Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduction and Targeted Therapy, 4*(1), 62. doi:10.1038/s41392-019-0095-0

Clark, S. J., Statham, A., Stirzaker, C., Molloy, P. L., & Frommer, M. (2006). DNA methylation: bisulphite modification and analysis. *Nature protocols, 1*(5), 2353.

Compere, S. J., & Palmiter, R. D. (1981). DNA methylation controls the inducibility of the mouse metallothionein-I gene lymphoid cells. *Cell, 25*(1), 233-240. doi:10.1016/0092-8674(81)90248-8

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications, 19*(4), 497-515. doi:10.1007/s10260-010-0142-z

Daniel, M., & Tollefsbol, T. O. (2015). Epigenetic linkage of aging, cancer and nutrition. *J Exp Biol, 218*(Pt 1), 59-70. doi:10.1242/jeb.107110

Deatherage, D. E., Potter, D., Yan, P. S., Huang, T. H. M., & Lin, S. (2009). Methylation analysis by microarray. *Methods in molecular biology (Clifton, N.J.), 556*, 117-139. doi:10.1007/978-1-60327-192-9_9

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development, 25*(10), 1010-1022. doi:10.1101/gad.2037511

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., & Fuks, F. (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics, 15*(6), 929-941. doi:10.1093/bib/bbt054

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics, 3*(6), 771-784. doi:10.2217/epi.11.105

Dong, Y., Zhao, H., Li, H., Li, X., & Yang, S. (2014). DNA methylation as an early diagnostic marker of cancer (Review). *Biomed Rep, 2*(3), 326-330. doi:10.3892/br.2014.237

Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B., & Nestor, C. E. (2020). No evidence for DNA N⁶-methyladenine in mammals. *Science Advances, 6*(12), eaay3335. doi:10.1126/sciadv.aay3335

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics, 11*(1), 587. doi:10.1186/1471-2105-11-587

Duncan, B. K., & Miller, J. H. (1980). Mutagenic deamination of cytosine residues in DNA. *Nature, 287*(5782), 560-561. doi:10.1038/287560a0

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research, 30*(1), 207-210. doi:10.1093/nar/30.1.207

Esteller, M. (2000). Epigenetic lesions causing genetic lesions in human cancer: promoter hypermethylation of DNA repair genes. *Eur J Cancer, 36*(18), 2294-2300. doi:10.1016/s0959-8049(00)00303-8

Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene, 21*(35), 5427-5440. doi:10.1038/sj.onc.1205600

Esteller, M. (2005). Aberrant DNA methylation as a cancer-inducing mechanism. *Annu Rev Pharmacol Toxicol, 45*, 629-656. doi:10.1146/annurev.pharmtox.45.120403.095832

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., . . . May, B. (2018). The reactome pathway knowledgebase. *Nucleic acids research, 46*(D1), D649-D655.

Ferrari, L., Vicenzi, M., Tarantini, L., Barretta, F., Sironi, S., Baccarelli, A. A., . . . Bollati, V. (2019). Effects of Physical Exercise on Endothelial Function and DNA Methylation. *International journal of environmental research and public health, 16*(14), 2530. doi:10.3390/ijerph16142530

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., . . . Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology, 15*(12), 503-503. doi:10.1186/s13059-014-0503-2

Fortin, J.-P., Triche, T. J., Jr., & Hansen, K. D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics (Oxford, England), 33*(4), 558-560. doi:10.1093/bioinformatics/btw691

Fraser, R., & Lin, C.-J. (2016). Epigenetic reprogramming of the zygote in mice and men: on your marks, get set, go! *Reproduction (Cambridge, England), 152*(6), R211-R222. doi:10.1530/REP-16-0376

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., . . . Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America, 89*(5), 1827-1831. doi:10.1073/pnas.89.5.1827

Gaine, M. E., Chatterjee, S., & Abel, T. (2018). Sleep Deprivation and the Epigenome. *Frontiers in neural circuits, 12*, 14-14. doi:10.3389/fncir.2018.00014

Gardiner-Garden, M., & Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology, 196*(2), 261-282. doi:https://doi.org/10.1016/0022-2836(87)90689-9

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., . . . Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome biology, 17*(1), 61. doi:10.1186/s13059-016-0926-z

Genome Reference Consortium. (2021). Human Genome Issues. Retrieved from https://www.ncbi.nlm.nih.gov/grc/human/issues?filters=type:gap&asm=GRCh37.p13,

Grönniger, E., Weber, B., Heil, O., Peters, N., Stäb, F., Wenck, H., . . . Lyko, F. (2010). Aging and chronic sun exposure cause distinct epigenetic changes in human skin. *PLoS genetics, 6*(5), e1000971-e1000971. doi:10.1371/journal.pgen.1000971

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX*. Retrieved from

Hannon, G. J. (2002). RNA interference. *Nature, 418*(6894), 244-251. doi:10.1038/418244a

Hansen, R. S., Wijmenga, C., Luo, P., Stanek, A. M., Canfield, T. K., Weemaes, C. M., & Gartler, S. M. (1999). The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences of the United States of America, 96*(25), 14412-14417. doi:10.1073/pnas.96.25.14412

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Searle, S. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research, 22*(9), 1760-1774.

Heijmans, B. T., Tobi, E. W., Stein, A. D., Putter, H., Blauw, G. J., Susser, E. S., . . . Lumey, L. H. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(44), 17046-17049. doi:10.1073/pnas.0806560105

Heyn, H., & Esteller, M. (2015). An Adenine Code for DNA: A Second Life for N6-Methyladenine. *Cell, 161*(4), 710-713. doi:10.1016/j.cell.2015.04.021

Heyn, H., Li, N., Ferreira, H. J., Moran, S., Pisano, D. G., Gomez, A., . . . Esteller, M. (2012). Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences, 109*(26), 10522. doi:10.1073/pnas.1120658109

Heyn, H., Vidal, E., Sayols, S., Sanchez-Mut, J. V., Moran, S., Medina, I., . . . Esteller, M. (2012). Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient. *Epigenetics, 7*(6), 542-550. doi:10.4161/epi.20523

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome biology, 14*(10), 3156. doi:10.1186/gb-2013-14-10-r115

Horvath, S., Pirazzini, C., Bacalini, M. G., Gentilini, D., Di Blasio, A. M., Delledonne, M., . . . Franceschi, C. (2015). Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. *Aging (Albany NY), 7*(12), 1159-1170. doi:10.18632/aging.100861

Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics, 19*(6), 371-384.

Hossain, K., Suzuki, T., Hasibuzzaman, M., Islam, M. S., Rahman, A., Paul, S. K., . . . Rahman, M. (2017). Chronic exposure to arsenic, LINE-1 hypomethylation, and blood pressure: a cross-sectional study in Bangladesh. *Environmental Health, 16*(1), 1-12.

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem, 175*(1), 315-332.

Houseman, E. A., Kim, S., Kelsey, K. T., & Wiencke, J. K. (2015). DNA Methylation in Whole Blood: Uses and Challenges. *Current Environmental Health Reports, 2*(2), 145-154. doi:10.1007/s40572-015-0050-3

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *IEEE Annals of the History of Computing, 9*(03), 90-95.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics, 5*(3), 299-314.

Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., . . . Bird, A. P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics, 6*(9), e1001134-e1001134. doi:10.1371/journal.pgen.1001134

Illumina Inc. (2010). GenomeStudio Methylation Module v1.8 User Guide. In (pp. 51).

Illumina Inc. (2020a). How Do Illumina Microarrays Work? Retrieved from https://www.illumina.com/science/technology/microarray.html

Illumina Inc. (2020b). Infinium Methylation Assay Overview. Retrieved from https://www.illumina.com/science/technology/microarray.html

Issa, J.-P. (2014). Aging and epigenetic drift: a vicious cycle. *The Journal of clinical investigation, 124*(1), 24-29. doi:10.1172/JCI69735

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet, 33 Suppl*, 245-254. doi:10.1038/ng1089

Johnson, A. A., Akman, K., Calimport, S. R. G., Wuttke, D., Stolzing, A., & de Magalhães, J. P. (2012). The role of DNA methylation in aging, rejuvenation, and age-related disease. *Rejuvenation research, 15*(5), 483-494. doi:10.1089/rej.2012.1324

Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., . . . Cupul-Uicab, L. A. (2012). 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental health perspectives, 120*(10), 1425-1431.

Kadayifci, F. Z., Zheng, S., & Pan, Y.-X. (2018). Molecular Mechanisms Underlying the Link between Diet and DNA Methylation. *International journal of molecular sciences, 19*(12), 4055. doi:10.3390/ijms19124055

Kane, A. E., & Sinclair, D. A. (2019). Epigenetic changes during aging and their reprogramming potential. *Critical reviews in biochemistry and molecular biology, 54*(1), 61-83. doi:10.1080/10409238.2019.1570075

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., & Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature, 429*(6994), 900-903. doi:10.1038/nature02633

Kanherkar, R. R., Bhatia-Dey, N., & Csoka, A. B. (2014). Epigenetics across the human lifespan. *Frontiers in cell and developmental biology, 2*, 49-49. doi:10.3389/fcell.2014.00049

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y., . . . Thomas, D. J. (2003). The UCSC genome browser database. *Nucleic acids research, 31*(1), 51-54.

Kiefer, J. C. (2007). Epigenetics in development. *Developmental Dynamics, 236*(4), 1144-1156. doi:10.1002/dvdy.21094

Kile, M. L., Baccarelli, A., Hoffman, E., Tarantini, L., Quamruzzaman, Q., Rahman, M., . . . Wright, R. O. (2012). Prenatal arsenic exposure and DNA methylation in maternal and umbilical cord blood leukocytes. *Environmental health perspectives, 120*(7), 1061-1066.

Kohli, R. M., & Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. *Nature, 502*(7472), 472-479. doi:10.1038/nature12750

Komaki, S., Shiwa, Y., Furukawa, R., Hachiya, T., Ohmomo, H., Otomo, R., . . . Shimizu, A. (2018). iMETHYL: an integrative database of human DNA methylation, gene expression, and genomic variation. *Human Genome Variation, 5*(1), 18008. doi:10.1038/hgv.2018.8

Krokan, H. E., & Bjørås, M. (2013). Base excision repair. *Cold Spring Harbor perspectives in biology, 5*(4), a012583-a012583. doi:10.1101/cshperspect.a012583

Lambrou, A., Baccarelli, A., Wright, R. O., Weisskopf, M., Bollati, V., Amarasiriwardena, C., . . . Schwartz, J. (2012). Arsenic exposure and DNA methylation among elderly men. *Epidemiology (Cambridge, Mass.), 23*(5), 668.

Larsen, F., Gundersen, G., Lopez, R., & Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics, 13*(4), 1095-1107. doi:10.1016/0888-7543(92)90024-m

Li, R., Liang, F., Li, M., Zou, D., Sun, S., Zhao, Y., . . . Zhang, Z. (2018). MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic acids research, 46*(D1), D288-D295. doi:10.1093/nar/gkx1139

Liu, L., Wylie, R. C., Andrews, L. G., & Tollefsbol, T. O. (2003). Aging, cancer and nutrition: the DNA methylation connection. *Mech Ageing Dev, 124*(10-12), 989-998. doi:10.1016/j.mad.2003.08.001

Lu, Y., Brommer, B., Tian, X., Krishnan, A., Meer, M., Wang, C., . . . Sinclair, D. A. (2020). Reprogramming to recover youthful epigenetic information and restore vision. *Nature, 588*(7836), 124-129. doi:10.1038/s41586-020-2975-4

Lucchesi, J. C., Kelly, W. G., & Panning, B. (2005). Chromatin remodeling in dosage compensation. *Annu Rev Genet, 39*, 615-651. doi:10.1146/annurev.genet.39.073003.094210

Luo, Y., Lu, X., & Xie, H. (2014). Dynamic Alu methylation during normal development, aging, and tumorigenesis. *BioMed research international, 2014*, 784706-784706. doi:10.1155/2014/784706

Lyko, F. (2018). The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics, 19*(2), 81-92. doi:10.1038/nrg.2017.80

LYON, M. F. (1972). X-CHROMOSOME INACTIVATION AND DEVELOPMENTAL PATTERNS IN MAMMALS. *Biological Reviews, 47*(1), 1-35. doi:https://doi.org/10.1111/j.1469-185X.1972.tb00969.x

Madrigano, J., Baccarelli, A. A., Mittleman, M. A., Sparrow, D., Vokonas, P. S., Tarantini, L., & Schwartz, J. (2012). Aging and epigenetics: Longitudinal changes in gene-specific DNA methylation. *Epigenetics, 7*(1), 63-70. doi:10.4161/epi.7.1.18749

Maierhofer, A., Flunkert, J., Oshima, J., Martin, G. M., Haaf, T., & Horvath, S. (2017). Accelerated epigenetic aging in Werner syndrome. *Aging (Albany NY), 9*(4), 1143-1152. doi:10.18632/aging.101217

Maksimovic, J., Gordon, L., & Oshlack, A. (2012). SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome biology, 13*(6), R44. doi:10.1186/gb-2012-13-6-r44

Marioni, R. E., Shah, S., McRae, A. F., Ritchie, S. J., Muniz-Terrera, G., Harris, S. E., . . . Deary, I. J. (2015). The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int J Epidemiol, 44*(4), 1388-1396. doi:10.1093/ije/dyu277

Martin, E. M., & Fry, R. C. (2018). Environmental Influences on the Epigenome: Exposure- Associated DNA Methylation in Human Populations. *Annual Review of Public Health, 39*(1), 309-333. doi:10.1146/annurev-publhealth-040617-014629

McKinney, W. (2010). *Data structures for statistical computing in python.* Paper presented at the Proceedings of the 9th Python in Science Conference.

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 38*(1), 23-38. doi:10.1038/npp.2012.112

Mossman, D., & Scott, R. J. (2006). Epimutations, inheritance and causes of aberrant DNA methylation in cancer. *Hereditary cancer in clinical practice, 4*(2), 75-80. doi:10.1186/1897-4287-4-2-75

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., & Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature, 393*(6683), 386-389. doi:10.1038/30764

Ni, P., Huang, N., Zhang, Z., Wang, D.-P., Liang, F., Miao, Y., . . . Wang, J. (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics (Oxford, England), 35*(22), 4586-4595. doi:10.1093/bioinformatics/btz276

Nicodemus-Johnson, J., & Sinnott, R. A. (2017). Fruit and Juice Epigenetic Signatures Are Associated with Independent Immunoregulatory Pathways. *Nutrients, 9*(7), 752. doi:10.3390/nu9070752

Noble, A. (2021). *The impact of the environment on DNA methylation in humans and zebrafish.* (Doctor of Philosophy). University of Canterbury,

Osborne, A. J., Pearson, J. F., Noble, A. J., Gemmell, N. J., Horwood, L. J., Boden, J. M., . . . Kennedy, M. A. (2020). Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort. *Translational Psychiatry, 10*(1), 114. doi:10.1038/s41398-020-0800-3

Partington, M. W., Robinson, H., Laing, S., & Turner, G. (1992). Mortality in the fragile X syndrome: Preliminary data. *American Journal of Medical Genetics, 43*(1-2), 120-123. doi:https://doi.org/10.1002/ajmg.1320430118

Paska, A. V., & Hudler, P. (2015). Aberrant methylation patterns in cancer: a clinical view. *Biochemia medica, 25*(2), 161-176. doi:10.11613/BM.2015.017

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london, 60*(359-367), 489-498.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Peluso, M. E., Munnia, A., Bollati, V., Srivatanakul, P., Jedpiyawongse, A., Sangrajrang, S., . . . Baccarelli, A. A. (2014). Aberrant methylation of hypermethylated-in-cancer-1 and exocyclic DNA adducts in tobacco smokers. *toxicological sciences, 137*(1), 47-54.

Peng, H., Zhu, Y., Goldberg, J., Vaccarino, V., & Zhao, J. (2019). DNA Methylation of Five Core Circadian Genes Jointly Contributes to Glucose Metabolism: A Gene-Set Analysis in Monozygotic Twins. *Frontiers in Genetics, 10*(329). doi:10.3389/fgene.2019.00329

Pérez, R. F., Santamarina, P., Tejedor, J. R., Urdinguio, R. G., Álvarez-Pitti, J., Redon, P., . . . Lurbe, E. (2019). Longitudinal genome-wide DNA methylation analysis uncovers persistent early-life DNA methylation changes. *Journal of Translational Medicine, 17*(1), 15. doi:10.1186/s12967-018-1751-9

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Breitwieser, F. P., Chang, Y.-C., . . . Salzberg, S. L. (2018). CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome biology, 19*(1), 208. doi:10.1186/s13059-018-1590-2

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., . . . Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome biology, 17*(1), 208-208. doi:10.1186/s13059-016-1066-1

Pombo, A., & Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology, 16*(4), 245-257. doi:10.1038/nrm3965

Rae, M. J., Butler, R. N., Campisi, J., de Grey, A. D. N. J., Finch, C. E., Gough, M., . . . Logan, B. J. (2010). The demographic and biomedical case for late-life interventions in aging. *Science translational medicine, 2*(40), 40cm21-40cm21. doi:10.1126/scitranslmed.3000822

Reik, W., & Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nat Rev Genet, 2*(1), 21-32. doi:10.1038/35047554

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., & Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics (Oxford, England), 23*(20), 2700-2707. doi:10.1093/bioinformatics/btm412

Robertson, K. D., & Wolffe, A. P. (2000). DNA methylation in health and disease. *Nature Reviews Genetics, 1*(1), 11-19. doi:10.1038/35049533

Rogers, M. A., Langbein, L., Winter, H., Ehmann, C., Praetzel, S., & Schweizer, J. (2002). Characterization of a first domain of human high glycine-tyrosine and high sulfur keratin-associated protein (KAP) genes on chromosome 21q22.1. *J Biol Chem, 277*(50), 48993-49002. doi:10.1074/jbc.M206422200

Roloff, T. C., Ropers, H. H., & Nuber, U. A. (2003). Comparative study of methyl-CpG-binding domain proteins. *BMC genomics, 4*(1), 1-1. doi:10.1186/1471-2164-4-1

Saghafinia, S., Mina, M., Riggi, N., Hanahan, D., & Ciriello, G. (2018). Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors. *Cell Reports, 25*(4), 1066-1080.e1068. doi:https://doi.org/10.1016/j.celrep.2018.09.082

Sailani, M. R., Halling, J. F., Møller, H. D., Lee, H., Plomgaard, P., Pilegaard, H., . . . Regenberg, B. (2019). Lifelong physical activity is associated with promoter hypomethylation of genes involved in metabolism, myogenesis, contractile properties and oxidative stress resistance in aged human skeletal muscle. *Scientific Reports, 9*(1), 3272. doi:10.1038/s41598-018-37895-8

Sala, C., Di Lena, P., Fernandes Durso, D., Prodi, A., Castellani, G., & Nardini, C. (2020). Evaluation of pre-processing on the meta-analysis of DNA methylation data from the Illumina HumanMethylation450 BeadChip platform. *PLOS ONE, 15*(3), e0229763. doi:10.1371/journal.pone.0229763

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences, 103*(5), 1412. doi:10.1073/pnas.0510310103

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609-612.

Sharp, A. J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S. B., Dupre, Y., & Antonarakis, S. E. (2011). DNA methylation profiles of human active and inactive X chromosomes. *Genome research, 21*(10), 1592-1600. doi:10.1101/gr.112680.110

Shenker, N. S., Ueland, P. M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., . . . Vineis, P. (2013). DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*, 712-716.

Smeets, E. E. J., Pelc, K., & Dan, B. (2012). Rett Syndrome. *Molecular syndromology, 2*(3-5), 113-127. doi:10.1159/000337637

Smith, Z. D., & Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics, 14*(3), 204-220. doi:10.1038/nrg3354

Solomon, O., MacIsaac, J., Quach, H., Tindula, G., Kobor, M. S., Huen, K., . . . Holland, N. (2018). Comparison of DNA methylation measured by Illumina 450K and EPIC BeadChips in blood of newborns and 14-year-old children. *Epigenetics, 13*(6), 655-664. doi:10.1080/15592294.2018.1497386

Soto-Ramírez, N., Arshad, S. H., Holloway, J. W., Zhang, H., Schauberger, E., Ewart, S., . . . Karmaus, W. (2013). The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics, 5*(1), 1. doi:10.1186/1868-7083-5-1

Sun, H., Zhou, H., Zhang, Y., Chen, J., Han, X., Huang, D., . . . Zhao, Y. (2018). Aberrant methylation of *FAT4* and *SOX11* in peripheral blood leukocytes and their association with gastric cancer risk. *Journal of Cancer, 9*(13), 2275-2283. doi:10.7150/jca.24797

Takeshima, H., Yamada, H., & Ushijima, T. (2019). Chapter 5 - Cancer Epigenetics: Aberrant DNA Methylation in Cancer Diagnosis and Treatment. In F. Dammacco & F. Silvestris (Eds.), *Oncogenomics* (pp. 65-76): Academic Press.

Talens, R. P., Christensen, K., Putter, H., Willemsen, G., Christiansen, L., Kremer, D., . . . Heijmans, B. T. (2012). Epigenetic variation during the adult lifespan: cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell, 11*(4), 694-703. doi:https://doi.org/10.1111/j.1474-9726.2012.00835.x

Thévenin, A., Ein-Dor, L., Ozery-Flato, M., & Shamir, R. (2014). Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic acids research, 42*(15), 9854-9861. doi:10.1093/nar/gku667

Thompson, R. F., Atzmon, G., Gheorghe, C., Liang, H. Q., Lowes, C., Greally, J. M., & Barzilai, N. (2010). Tissue-specific dysregulation of DNA methylation in aging. *Aging Cell, 9*(4), 506-518. doi:10.1111/j.1474-9726.2010.00577.x

Thompson, R. F., Atzmon, G., Gheorghe, C., Liang, H. Q., Lowes, C., Greally, J. M., & Barzilai, N. (2010). Tissue-specific dysregulation of DNA methylation in aging. *Aging Cell, 9*(4), 506-518. doi:10.1111/j.1474-9726.2010.00577.x

Touleimat, N., & Tost, J. (2012). Complete pipeline for Infinium(®) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics, 4*(3), 325-341. doi:10.2217/epi.12.21

Trerotola, M., Relli, V., Simeone, P., & Alberti, S. (2015). Epigenetic inheritance and the missing heritability. *Human genomics, 9*(1), 17-17. doi:10.1186/s40246-015-0041-3

Triche, T. J., Jr., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., & Siegmund, K. D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic acids research, 41*(7), e90-e90. doi:10.1093/nar/gkt090

Trowbridge, J. J., Snow, J. W., Kim, J., & Orkin, S. H. (2009). DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. *Cell Stem Cell, 5*(4), 442-449. doi:10.1016/j.stem.2009.08.016

Tse, J. W. T., Jenkins, L. J., Chionh, F., & Mariadason, J. M. (2017). Aberrant DNA Methylation in Colorectal Cancer: What Should We Target? *Trends in Cancer, 3*(10), 698-712. doi:10.1016/j.trecan.2017.08.003

Ulrey, C. L., Liu, L., Andrews, L. G., & Tollefsbol, T. O. (2005). The impact of metabolism on DNA methylation. *Hum Mol Genet, 14 Spec No 1*, R139-147. doi:10.1093/hmg/ddi100

University of California Santa Cruz. (2020). Cpg Islands Cpg Islands Track Settings. Retrieved from http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=cpgIslandExt

Unnikrishnan, A., Hadad, N., Masser, D. R., Jackson, J., Freeman, W. M., & Richardson, A. (2018). Revisiting the genomic hypomethylation hypothesis of aging. *Annals of the New York Academy of Sciences, 1418*(1), 69-79. doi:10.1111/nyas.13533

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering, 13*(2), 22-30.

van Roekel E. H., Dugué, P. A., Jung, C. H., Joo, J. E., Makalic, E., Wong, E. E. M., . . . Milne, R. L. (2019). Physical Activity, Television Viewing Time, and DNA Methylation in Peripheral Blood. *Med Sci Sports Exerc, 51*(3), 490-498. doi:10.1249/mss.0000000000001827

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*: Centrum voor Wiskunde en Informatica Amsterdam.

Veenendaal, M. V., Painter, R. C., de Rooij, S. R., Bossuyt, P. M., van der Post, J. A., Gluckman, P. D., . . . Roseboom, T. J. (2013). Transgenerational effects of prenatal exposure to the 1944-45 Dutch famine. *Bjog, 120*(5), 548-553. doi:10.1111/1471-0528.12136

Veenstra, J., Kalsbeek, A., Koster, K., Ryder, N., Bos, A., Huisman, J., . . . Tintle, N. L. (2018). Epigenome wide association study of SNP-CpG interactions on changes in triglyceride levels after pharmaceutical intervention: a GAW20 analysis. *BMC proceedings, 12*(Suppl 9), 58-58. doi:10.1186/s12919-018-0144-7

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., . . . Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods, 17*(3), 261-272.

Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., . . . Liu, J. (2014). Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell, 157*(4), 979-991. doi:10.1016/j.cell.2014.04.017

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., . . . Chen, W. (2015). A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics, 10*(7), 662-669. doi:10.1080/15592294.2015.1057384

Winkler, G. S. (2010). The mammalian anti-proliferative BTG/Tob protein family. *J Cell Physiol, 222*(1), 66-72. doi:10.1002/jcp.21919

Wu, M. C., Joubert, B. R., Kuan, P.-f., Håberg, S. E., Nystad, W., Peddada, S. D., & London, S. J. (2014). A systematic assessment of normalization approaches for the Infinium 450K methylation platform. *Epigenetics, 9*(2), 318-329. doi:10.4161/epi.27119

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., . . . Xiao, A. Z. (2016). DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature, 532*(7599), 329-333. doi:10.1038/nature17640

Wu, X., & Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet, 18*(9), 517-534. doi:10.1038/nrg.2017.33

Xiao, C. L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., . . . Yan, G. R. (2018). N(6)-Methyladenine DNA Modification in the Human Genome. *Mol Cell, 71*(2), 306-318.e307. doi:10.1016/j.molcel.2018.06.015

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., . . . Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature, 500*(7463), 477-481. doi:10.1038/nature12433